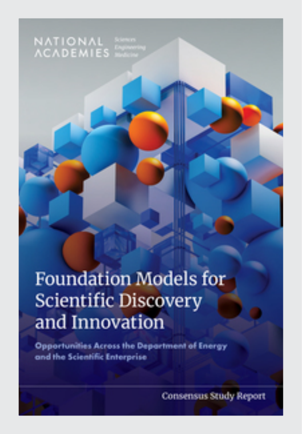
NATIONAL ACADEMIES Sciences Engineering Medicine
Foundation Models for Scientific Discovery and Innovation
Opportunities Across the Department of Energy and the Scientific Enterprise
Consensus Study Report

Foundation Models for Scientific Discovery and Innovation: Opportunities Across the Department of Energy and the Scientific Enterprise (2025)

DETAILS

98 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-99500-9 | DOI 10.17226/29212

CONTRIBUTORS

Committee on Foundation Models for Scientific Discovery and Innovation; Board on Mathematical Sciences and Analytics; Division on Engineering and Physical Sciences; National Academies of Sciences, Engineering, and Medicine

SUGGESTED CITATION

National Academies of Sciences, Engineering, and Medicine. 2025. *Foundation Models for Scientific Discovery and Innovation: Opportunities Across the Department of Energy and the Scientific Enterprise*. Washington, DC: The National Academies Press. https://doi.org/10.17226/29212.

NATIONAL ACADEMIES *Sciences Engineering Medicine*

NATIONAL ACADEMIES PRESS
Washington, DC

Foundation Models for Scientific Discovery and Innovation

Opportunities Across the Department of Energy and the Scientific Enterprise

Committee on Foundation Models for Scientific Discovery and Innovation

Board on Mathematical Sciences and Analytics

Division on Engineering and Physical Sciences

Consensus Study Report

NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

This project was supported by Award DE-SC0002220 between the National Academy of Sciences and the Department of Energy. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

International Standard Book Number-13: 978-0-309-99500-9
Digital Object Identifier: https://doi.org/10.17226/29212

This publication is available from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242; https://nap.nationalacademies.org.

The manufacturer's authorized representative in the European Union for product safety is Authorised Rep Compliance Ltd., Ground Floor, 71 Lower Baggot Street, Dublin D02 P593 Ireland; www.arccompliance.com.

Printed in the United States of America.

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2025. *Foundation Models for Scientific Discovery and Innovation: Opportunities Across the Department of Energy and the Scientific Enterprise*. Washington, DC: National Academies Press. https://doi.org/10.17226/29212.

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. Tsu-Jae Liu is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.nationalacademies.org**.

Consensus Study Reports published by the National Academies of Sciences, Engineering, and Medicine document the evidence-based consensus on the study's statement of task by an authoring committee of experts. Reports typically include findings, conclusions, and recommendations based on information gathered by the committee and the committee's deliberations. Each report has been subjected to a rigorous and independent peer-review process and it represents the position of the National Academies on the statement of task.

Proceedings published by the National Academies of Sciences, Engineering, and Medicine chronicle the presentations and discussions at a workshop, symposium, or other event convened by the National Academies. The statements and opinions contained in proceedings are those of the participants and are not endorsed by other participants, the planning committee, or the National Academies.

Rapid Expert Consultations published by the National Academies of Sciences, Engineering, and Medicine are authored by subject-matter experts on narrowly focused topics that can be supported by a body of evidence. The discussions contained in rapid expert consultations are considered those of the authors and do not contain policy recommendations. Rapid expert consultations are reviewed by the institution before release.

For information about other products and activities of the National Academies, please visit www.nationalacademies.org/about/whatwedo.

COMMITTEE ON FOUNDATION MODELS FOR SCIENTIFIC DISCOVERY AND INNOVATION

DONA L. CRAWFORD, Lawrence Livermore National Laboratory (retired), *Chair*
SYED BAHAUDDIN ALAM, University of Illinois Urbana-Champaign
MARTA D'ELIA, Atomic Machines and Stanford University
KRISHNA GARIKIPATI, University of Southern California
SHIRLEY HO, Flatiron Institute
SCOTT H. HOLAN, University of Missouri
MICHAEL KEARNS (NAS), University of Pennsylvania
PETROS KOUMOUTSAKOS (NAE), Harvard University
BRIAN KULIS, Boston University
DANIEL I. MEIRON, California Institute of Technology
NATHANIEL TRASK, University of Pennsylvania

Study Staff

BLAKE REICHMUTH, Associate Program Officer, Board on Mathematical Sciences and Analytics (BMSA), *Study Director*
MICHELLE SCHWALBE, Director, BMSA and National Materials and Manufacturing Board (NMMB)
ERIK SVEDBERG, Deputy Director, BMSA and NMMB
JON EISENBERG, Director, Computer Science and Telecommunications Board (CSTB)
THƠ H. NGUYỄN, Senior Program Officer, CSTB
SAM KORETSKY, Research Associate, BMSA
HEATHER LOZOWSKI, Senior Finance Business Partner
JOE PALMER, Senior Project Assistant, BMSA

BOARD ON MATHEMATICAL SCIENCES AND ANALYTICS

JILL C. PIPHER, Brown University, *Co-Chair*
KAREN E. WILLCOX, Oden Institute for Computational Engineering and Sciences, *Co-Chair*
BRENT APPLEBY, Draper Laboratory
HÉLÈNE BARCELO, Mathematical Sciences Research Institute
RUSSEL E. CAFLISCH, New York University
A. ROBERT CALDERBANK, Duke University
DUANE COOPER, Morehouse College
SARA Y. DEL VALLE, Los Alamos National Laboratory
RONALD D. FRICKER, JR., Virginia Polytechnic Institute and State University
SKIP GARIBALDI, Center for Communications Research, Institute for Defense Analyses
JULIE SIMMONS IVY, University of Michigan
PETROS KOUMOUTSAKOS (NAE), Harvard University
RACHEL KUSKE, Georgia Institute of Technology
YANN A. LeCUN, Facebook
AARTI SINGH, Carnegie Mellon University
KAREN E. SMITH, University of Michigan
BRANI VIDAKOVIC, Texas A&M University
JUDY WALKER, University of Nebraska–Lincoln

Study Staff

MICHELLE K. SCHWALBE, Director
ERIK SVEDBERG, Deputy Director
HEATHER LOZOWSKI, Senior Finance Business Partner
BRITTANY SEGUNDO, Program Officer
BLAKE REICHMUTH, Associate Program Officer
SAMANTHA KORETSKY, Research Assistant
JOE PALMER, Senior Project Assistant

COMPUTER SCIENCE AND TELECOMMUNICATIONS BOARD

LAURA M. HAAS, University of Massachusetts Amherst, *Chair*
DAVID DANKS, University of California, San Diego
ECE KAMAR, Microsoft Research
JAMES F. KUROSE, University of Massachusetts Amherst (emeritus)
DAVID LUEBKE, NVIDIA
DAWN C. MEYERRIECKS, The MITRE Corporation
WILLIAM L. SCHERLIS, Carnegie Mellon University
HENNING G. SCHULZRINNE, Columbia University
NAMBIRAJAN SESHADRI, University of California, San Diego
KENNETH E. WASHINGTON, Medtronic
JOHN L. MANFERDELLI, Independent Consultant (ex officio member)

Study Staff

JON EISENBERG, Senior Board Director
AARYA SHRESTHA, Senior Financial Business Partner
THƠ H. NGUYỄN, Senior Program Officer
GABRIELLE RISICA, Program Officer
SHANAE BRADLY, Administrative Coordinator

Reviewers

This Consensus Study Report was reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide candid and critical comments that will assist the National Academies of Sciences, Engineering, and Medicine in making each published report as sound as possible and to ensure that it meets the institutional standards for quality, objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process.

We thank the following individuals for their review of this report:

FRANCIS JOSEPH ALEXANDER, Argonne National Laboratory
ANDREA BERTOZZI (NAS), University of California, Los Angeles
BRIAN SCOTT CAFFO, Johns Hopkins University
WEI CHEN (NAE), Northwestern University
KARTHIK DURAISAMY, University of Michigan
JANE PINELIS, Johns Hopkins University Applied Physics Laboratory
ROBERT SCHAPIRE (NAS/NAE), Microsoft
FRED STREITZ, Lawrence Livermore National Laboratory (retired)

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations of this report nor did they see the final draft before its release. The review of this report was overseen by **REBECCA WILLET**, University of Chicago, and **ROBERT F. SPROULL**, University of Massachusetts Amherst. They were responsible for making certain that an independent examination of this report was carried out in accordance with the standards of the National Academies and that all review comments were carefully considered. Responsibility for the final content rests entirely with the authoring committee and the National Academies.

Acknowledgments

We are grateful to the many scholars and leaders who contributed their time and expertise to the committee's information-gathering efforts. Special thanks go to all the speakers who briefed the committee: William Collins, Lawrence Berkeley National Laboratory; Kevin Dixon, Sandia National Laboratories; Omar Ghattas, University of Texas at Austin; Hendrik Hamann, Brookhaven National Laboratory; Tzanio Kolev, Lawrence Livermore National Laboratory; Rémi Lam, Google DeepMind; Earl Lawrence, Los Alamos National Laboratory; Michael Mahoney, Lawrence Berkeley National Laboratory; Vivek Natarajan, Google DeepMind; Sabastian Nowozin, Google DeepMind; Chris Ritter, Idaho National Laboratory; Kelly Rose, National Energy Technology Laboratory; Brian Spears, Lawrence Livermore National Laboratory; Ann Speed, Sandia National Laboratories; Rick Stevens, Argonne National Laboratory; Georgia Tourassi, Oak Ridge National Laboratory; James Warren, National Institute of Standards and Technology; and Bin Yu, University of California, Berkeley.

Additionally, we appreciate the collaborative efforts of every member of the staff team. This report was the product of the committee's thoughtful deliberation and dedication on the topic. We extend our gratitude to the broader community for their engagement with this project.

Contents

APPENDIXES

Preface

Foundation models represent a potentially transformative technology for progressing scientific discovery and innovation. However, their rapid adoption has raised questions and concerns about their reliability, validity, and reproducibility. In 2024, the Department of Energy (DOE) requested that the National Academies of Sciences, Engineering, and Medicine conduct a study to consider current foundation models' capabilities, and future possibilities and challenges.

The National Academies established the Committee on Foundation Models for Scientific Discovery and Innovation to conduct this study. The study compares foundation models with more traditional computational methods, addresses exemplar use cases of foundation models, specifies strategic considerations, and outlines challenges for the development and use of foundation models. The full statement of the committee's task is shown in Appendix A.

The committee met in person in March 2025 and met virtually 15 times to receive briefings from experts and stakeholders (for a list of presentations, see Appendix B), review relevant reports and technical literature, deliberate, and develop this report.

The committee is grateful for the support of DOE's Office of Science, Office of Biological and Environmental Research, and National Nuclear Security Administration. The committee also extends its sincere thanks to the following National Academies' staff for their assistance throughout the study: Blake Reichmuth, Thơ Nguyễn, Erik Svedberg, Sam Koretsky, Jon Eisenberg, and Michelle Schwalbe.

Dona Crawford, *Chair*
Committee on Foundation Models for Scientific Discovery and Innovation
October 2025

Summary

There is significant interest in the development and application of foundation models for scientific discovery. Foundation models possess the capacity to generate outputs or findings and discern patterns within extensive data sets with data volumes that are considered overwhelming for classical modes of inquiry. Efforts are under way to use these models to accelerate various aspects of scientific workflows (including streamlining literature reviews, planning experiments, data analysis, and code development) and generating novel findings and hypotheses that can then spur further research directions. However, significant challenges remain in the effective use of these models in scientific applications, including issues with flawed or limited training data and limited verification, validation, and uncertainty quantification capabilities.

This report of the Committee on Foundation Models for Scientific Discovery and Innovation explores many of these opportunities and challenges and describes key gaps and potential future directions. This report explores use of foundation models independently and cooperatively with traditional modeling, exemplar use cases of foundation models, and challenges associated with the use of foundation models. While much of this report applies broadly to the use of foundation models for scientific discovery, the conversations are specifically focused on strategic considerations and directions for the Department of Energy (DOE) and its unique mission.

FOUNDATION MODELS AND TRADITIONAL MODELING

The current definition of foundation models varies across communities. This study uses the following definition:

> Today, foundation models are large-scale neural networks trained on vast amounts of heterogeneous data with the capability of learning new representations via fine-tuning on additional data. They represent a shift from traditional artificial intelligence (AI) systems designed for specific tasks. They possess the capacity to generate findings and discern patterns within extensive data sets with data volumes that exceed by orders of magnitude the computing and storage capacities of traditional solvers and even previous machine learning models.

Some of the key characteristics defining foundation models include massive scale, self-supervised pretraining, adaptability, emergent capabilities, ability to work in multiple modalities and be task agnostic, and a multipurpose architecture. These characteristics position foundation models as a potential paradigm shift for scientific research.

Despite the emergence of foundation models, traditional modeling (large-scale computational science solvers as well as statistical models) often retains critical advantages, particularly in interpretability, reliability, and strict adherence to physical laws. The fusion of traditional modeling approaches with foundation models is a promising direction.

> *Conclusion 2-1: Integrating traditional models with foundation models is proving to be increasingly powerful and has significant potential to advance computational findings in the physical sciences. These hybrid methods can be viewed as algorithmic alloys that can leverage the physical interpretability and structures of classical computational approaches alongside the data-driven adaptability of foundation models. This fusion enables the modeling of complex multiphysics, multiscale, and partially observed (understood) systems that challenge traditional approaches both computationally and mathematically.*

The fusion of foundation models with traditional numerical methods represents more than a computational advance; it constitutes a paradigm shift in the conduct of scientific discovery.

> **Recommendation 2-1: The Department of Energy (DOE) should invest in foundation model development, particularly in areas of strategic importance to DOE, including areas where DOE already has advantages leveraging its unique strengths in those domains. DOE should also prioritize the hybridization of foundation models and traditional modeling. Such hybrid modeling strategies can fuse the physical interpretability and robustness of classical solvers with the efficiency and learning capabilities of foundation models, particularly in multiscale, multiphysics applications where traditional approaches have limitations in capturing the heterogeneity,**

complexity, and dynamics of the physical system. DOE should not, however, abandon its expertise in numerical and computational methods and should continue investing strategically in software and infrastructure.

Although traditional modeling remains superior today in terms of interpretability and adherence to physical laws, integrating it with foundation models offers powerful new capabilities. These hybrid approaches enable better modeling of complex systems, and DOE should prioritize the use of these integrated methods.

EXEMPLAR USE CASES OF FOUNDATION MODELS

In framing potential DOE efforts in foundation models, the strategic focus remains a subject of debate: how best to balance the department's broad application space, navigate the trade-offs between leveraging past industry advancements and addressing the unique national security imperatives of DOE, and ensure responsible stewardship of taxpayer resources. A primary concern is that DOE cannot compete with the head start in technology maturation and large market share currently held by large companies, such as Microsoft and Google, that back efforts with large investments (both financially and with workforce).

Conclusion 3-1: Commercial industry has driven rapid progress in developing large language model–based foundation models, yielding a robust ecosystem of tools and capabilities. As demonstrated by, for example, the collaboration between Los Alamos National Laboratory and OpenAI, DOE can leverage these industry advances and findings as it develops foundation models for science and conducts coordinated DOE-wide assessments to identify appropriate opportunities.

This raises the fundamental question of whether DOE should be competing at all in the foundation model space and, if it does, whether it should focus on collaborations with industry or focus on complementary space where DOE's unique mission lies. The committee believes that DOE needs to develop these tools internally *in addition* to the private sector's development because the needs of the government, whether for national security or continued scientific preeminence, will not be met by private interests. The two endeavors (private and public) do not compete—they complement each other. Despite the mismatch in funding compared to industry leaders, DOE holds clear strategic advantages in several areas.

Conclusion 3-2: DOE retains clear strategic advantages in five areas: (1) a world-class scientific workforce in computational science; (2) access to large-scale, science-focused, and experimental computing *hardware; (3) stewardship of unique experimental facilities and open and controlled or classified scientific data; (4) capability to tackle long-term, high-risk, high-reward scientific problems; and (5) access to unique scientific data that may not be easily reproduced and which can be expanded as synthetic data may be necessary for training future foundation models.*

With DOE's advantages and foundation model capabilities in mind, the committee directed a series of recommendations addressing the potential role DOE can play in their development and implementation.

Keeping humans in the loop is important for foundation models for a number of reasons. These include addressing accountability and oversight, error detection and correction, interpretability and trust, and contextual judgment. The human counterpart can help determine the suitability and reliability of the foundation model. This outlines the following conclusion and recommendation from the committee regarding the importance of including humans in the foundation model processes.

Conclusion 3-3: While AI systems can exceed human performance in many ways, they can also fail in ways a human likely never would. For this reason, the qualification of foundation models will be necessary for decision making and prediction in the presence of uncertainty.

Recommendation 3-1: The Department of Energy (DOE) should study and develop the fusion of artificial intelligence (AI) and human capabilities. At present, AI systems handle the repetitive, manual, or routine tasks, and are starting to show abilities to reason. As AI becomes more capable, deep analysis and strategy recommendations become feasible, but humans should maintain oversight and validation, particularly for qualification and other aspects of DOE's mission.

Agentic AI has surged as a means of using large language models (LLMs) to launch external agents to explore hypotheses or improve or verify responses. There is a unique opportunity for DOE to explore these capabilities. Such capabilities may, for example, expose automatic differentiation "hooks" in their open-source libraries; to train foundation models, a software interface must expose the computational graph of a machine learning library to adjoint calculations in a scientific code, allowing the seamless backpropagation of gradients between the two codes. The majority of DOE codes are written in Fortran, C, or C++ and do not expose the necessary computational graph to pass adjoint information. If such hooks or interfaces were exposed, LLMs would be able to couple directly to production codes, integrating robust numerical prediction into the training process. This would allow LLMs to both perform simulation and calculate loss function, enabling complete end-to-end training with DOE's reliable and mature physics-based simulators. For example, a text prompt (e.g., "Why is the drag high on this vehicle?") could directly evaluate sensitivities to components of a scientific simulator (e.g., "The mesh facets of the tail fins are in a high-shear layer") rather than attempting to glean answers through text in scientific reports. By a similar process, DOE could apply LLMs and foundation models to help operate user facilities, leading to autonomous "self-driving laboratories."

> **Recommendation 3-2: The Department of Energy should evaluate the capabilities and risks of agentic artificial intelligence (AI) systems for its core applications. In particular, the committee advocates exploring agentic AI for developing autonomous laboratories for scientific discovery, decision making, and action planning for high-stakes applications.**

With the rapid development of foundation models and other AI systems, there is additional potential for security risks from these systems. The adversarial use of foundation models poses security risks in two main ways. First, attackers could target the model itself to subvert its function or steal intellectual property through methods such as *Prompt Injection* (jailbreaking), *Data Poisoning*, and *Model Stealing*. Second, adversaries could leverage foundation models as weapons to accelerate traditional cybercrime, enabling the mass production of highly effective phishing and deepfakes, lowering the barrier to writing malicious code, and introducing new supply chain vulnerabilities when models are integrated with external systems.

There needs to be the development of processes to verify that foundation models are reliable, safe, and trustworthy throughout their life cycles. Additional measures should also be developed to protect against adversarial applications of foundation models. These could be assisted by proactive cybersecurity strategies such as red teaming, where real-world attacks are simulated to help identify and address security weaknesses. The committee therefore states the following recommendation.

> **Recommendation 3-3: To address potential security risks arising from the adversarial use of foundation models, the Department of Energy should explore strategies for artificial intelligence assurance, red teaming, and development of countermeasures.**

Although industry leaders may have a head start with foundation models, there is value for DOE to focus on areas where it holds strategic advantages. Using these capabilities to help develop and direct foundation models can help to solidify DOE's place in foundation models for scientific discovery and innovation leadership.

STRATEGIC CONSIDERATIONS AND DIRECTIONS FOR THE DEPARTMENT OF ENERGY FOUNDATION MODELS

The national laboratories hold deep institutional expertise, embedded in their workforce, legacy data sets, and extensive experimental and modeling infrastructure. Yet the sheer scale of the DOE system, characterized by siloed specialized knowledge and the complexity of coordinating a large, distributed workforce, can be fundamentally misaligned with the speed and flexibility required for rapid decision making.

DOE invested early in material informatics and high-throughput experimental data curation campaigns to build unique access to data sets, through the Material Genome Initiative and other efforts. By combining advanced foundation models, high-performance computing, and curated experimental data, materials informatics can dramatically reduce the search space for viable material substitutes or processes. This is an example of a DOE effort to advance an aspect of computational science and how such research and development leads to important new capabilities.

Conclusion 4-1: Many DOE missions demand rapid analysis and decision making under urgent national security or economic constraints. Although the national laboratories hold deep institutional expertise—embedded in their workforce, legacy data sets, and extensive experimental and modeling infrastructure—the sheer scale of the DOE system, characterized by siloed specialized knowledge and the complexity of coordinating a large, distributed workforce, can be misaligned with the agility required for decisive action. Development of foundation models for this purpose poses a unique opportunity to address rapid analysis and decision making.

Recommendation 4-1: The Department of Energy should explore the use of foundation models to accelerate situational understanding by unifying dispersed, siloed, and diverse multimodal data sources as input to decision-making frameworks across heterogeneous environments.

Additionally, the needs of a DOE foundation model arguably pose more stringent requirements than in academic/industrial settings. For stockpile stewardship, simulation of critical components has matured over decades to the point that simulations calibrated by extensive testing are viewed as capable of replacing full-scale, experiment-based design. This outlines more opportunities for DOE.

Conclusion 4-2: DOE is uniquely positioned to shape the future of AI-driven science. Material informatics and near-autonomous scientific platforms highlight the power of combining curated experimental data, simulation, and advanced AI to accelerate discovery. Federated comput-

ing and facility integration extend this vision by enabling distributed use of DOE's infrastructure.

The curation and integration of specialized knowledge coupled with emerging multimodal and agentic AI approaches underscore the importance of preserving expertise, reasoning across diverse scientific data streams, and directly linking foundation models to DOE's mature simulation ecosystem.

Recommendation 4-2: The Department of Energy should both modernize existing infrastructure and invest in new infrastructure to generate, curate, and facilitate the large data corpus necessary to build a scientific foundation model, including simulations to create data, high-throughput and/or autonomous experimental facilities, and facilities to host data. Additionally, they should create interfaces (e.g., agentic, retrieval-augmented generation tools) through which large foundation models may easily access these sources. A successful strategy will provide holistic access to multimodal or heterogeneous infrastructure across the entire DOE complex, mitigating the "stovepiping" of assets between different laboratories or departments.

A strength of DOE is its ability to retain scientific talent, which should be reinforced with AI expertise as well. The success of any DOE-wide foundation model initiative depends entirely on attracting and retaining top AI talent, including overcoming the hurdle of slow funding cycles. However, DOE currently has excellent infrastructure and expertise as well as well-defined, mission-driven research.

Conclusion 4-3: DOE struggles to compete with the private sector for AI talent due to lower salaries and slow, traditional funding cycles. However, DOE's unique strengths, such as its mission-driven work, long-term career paths, and powerful supercomputing infrastructure, can be leveraged to attract talent. Building a strong academic pipeline through closer collaboration with universities is also essential for its long-term success.

Recommendation 4-3: To maintain a top-tier workforce, the Department of Energy (DOE) should design leadership-scale scientific research programs and provide staff with opportunities to rapidly adapt to a quickly evolving technological landscape. To attract early-career scientists, DOE should be perceived as the best place to become a leader in scientific machine learning; although industry may lead in large language model space, the unique access to state-of-the-art science can attract top talent. To be competitive with large-scale development efforts in industry, it is important to avoid

fracturing of scientists' time and attention. We recommend that DOE should create mechanisms by which medium through large teams can mount coordinated, focused efforts targeting mission-critical developments in fundamental research into, and applications of, foundation models for science.

One of DOE's major strengths is its data collection and generation capabilities. Leaning on this strength can be beneficial to the development and use of foundation models for scientific discovery and innovation. To expedite this potential, the data generated needs to be readable and usable by both the human users and the foundation models. This will help enhance operational efficiency and productivity and boost communication and collaboration. The standardization of data can help make DOE data readable and usable.

Conclusion 4-4: Although DOE curates many high-value data sets of value for construction of foundation models, they are typically developed in an ad hoc manner with heterogeneous file formats and data curation strategies that currently pose a barrier to high-throughput processing of data. Foundation models present a unique opportunity to address this issue.

Recommendation 4-4: To increase the success of future foundation models for science, the Department of Energy should invest in large-scale data user facilities (classified and unclassified), leveraged by artificial intelligence's growing capability to interpret heterogeneous scientific data, similar to the successes experienced with previous investments in supercomputers and open-source scientific computing libraries.

FOUNDATION MODEL CHALLENGES

Applying foundation models within DOE missions presents a multilayered set of scientific and operational challenges. These models, which emerged from success in domains such as natural language processing and vision, struggle to transfer directly into computational science workflows that demand physical fidelity, mesh-aware representations, and scalable performance across problems involving multiscale and multiphysics described by partial differential equations.

Verification, validation, and uncertainty quantification (VVUQ) are essential components of trustworthy scientific computing, ensuring that models are mathematically sound (verification), accurately represent the real-world systems they simulate (validation), and provide a clear understanding of uncertainties in their predictions (uncertainty quantification). These practices are well established in traditional modeling and simulation but are not yet adequately developed or standardized, particularly for foundation models. AI models often operate as black boxes, lacking transparency in how outputs are generated or how reliable they are under different conditions. Establishing VVUQ standards for foundation models is critical to ensure that these systems can be safely and effectively used in scientific discovery.

Conclusion 5-1: VVUQ methods analogous to those for traditional computational modeling do not exist for, or map directly onto, foundation models.

Conclusion 5-2: VVUQ, interpretability, and reproducibility are critical for establishing and maintaining trust in systems that are inherently complex, opaque, and increasingly deployed in high-stakes situations. Integration of VVUQ into foundation models would lead to increasing their trustworthiness, reliability, and fit for purpose, which is essential for future scientific discovery and innovation.

Recommendation 5-1: The Department of Energy (DOE) should lead the development of verification, validation, and uncertainty quantification frameworks tailored to foundation models, with built-in support for physical consistency, structured uncertainty quantification, and reproducible benchmarking in DOE-relevant settings.

There have been successes in validating model outputs with experimental data, as the data provide a real-world benchmark against which the models' accuracy can be determined. Without high-quality and robust experimental data, it is difficult to determine if a model's predictions are valid or merely artifacts of its assumptions or training data. This is especially important for foundation models and hybrid models, which may generalize well in theory but fail under specific conditions or in untested regimes. Therefore, the committee states the following conclusion and recommendation.

Conclusion 5-3: Foundation models for science will demand more and different physical experiments to validate the veracity of the AI predictions. Empirical grounding ensures that foundation model outputs reflect physical laws and domain-specific behavior. This is especially critical in high-stake DOE applications, where simulations alone cannot guarantee correctness, and where physical experiments provide the only definitive test of predictive validity.

Recommendation 5-2: In line with Recommendation 4-2, the Department of Energy should place high priority on data collection efforts to support reproducible foundation model training and validation, analogous to traditional efforts in verification, validation, and uncertainty quantification.

DOE is in a unique position for the development and use of foundation models for scientific discovery. They are leaders and have the capacity to tackle long-term, high-risk, high-reward scientific problems. An issue currently with foundation models for science is the lack of standards for development and use. Using these key resources, DOE can be contributing to the development and establishment of these standards for foundation models. Having concrete standards ensures compatibility and interoperability and improves reliability of a system.

Recommendation 5-3: The Department of Energy should establish and enforce standardized protocols and develop benchmarks for training, documenting, and reproducing foundation models for science and should participate in defining software standards, addressing randomness, hardware variability, and data access across its laboratories and high-performance computing infrastructure.

Although many of the technical challenges associated with foundation models can be addressed through internal research and development, deployment at DOE scale will increasingly involve external partnerships. Collaboration with industry introduces more constraints. Proprietary model weights, restricted data access, and closed-source infrastructure often prevent rigorous VVUQ and reproducibility practices, especially when security, transparency, or auditability is required. Collaboration with industry introduces more constraints. Proprietary model weights, restricted data access, and closed-source infrastructure often prevent rigorous VVUQ and reproducibility practices, especially when security, transparency, or auditability is required. These collaborations demand careful planning and coordination to bridge institutional differences in mission, priorities, and operational practices, particularly in areas such as contracting mechanisms, responsible AI standards, intellectual property frameworks, data-sharing protocols, and alignment on VVUQ expectations.

Conclusion 5-4: Partnering of DOE laboratories with industry on foundation models will require deliberate effort, including flexible contracting mechanisms, clear intellectual property agreements, data-sharing processes, aligning on VVUQ approaches, responsible AI practices, and a shared understanding of respective missions, objectives, and constraints.

Recommendation 5-4: The Department of Energy should deliberately pursue partnerships with industry and academia to address national mission goals, governed by flexible contracts, responsible artificial intelligence standards, and alignment on reproducibility, verification, validation, and uncertainty quantification approaches and data sharing.

1

Introduction

This chapter presents the rationale for the study, including any directive that led to its initiation. The statement of task for the study, the committee's interpretation of elements of the statement of task, and the structure of the report are also described. The chapter also points to potential reasons why the Department of Energy (DOE) is furthering the development and use of artificial intelligence (AI) models, such as foundation models. The study charge and the committee's interpretation of its key elements are then discussed, followed by a review of the report's structure in fulfillment of the study charge.

SIGNIFICANCE OF FOUNDATION MODELS

Foundation models are typically large-scale neural networks trained on vast amounts of heterogeneous data *with the capability of learning new representations via fine-tuning on additional data*. They represent a departure from traditional AI systems designed for specific tasks. They can be standalone systems or can be used as a "base" for many other applications (see Figure 1-1). Today, the most prominent foundation models are large language models (LLMs) trained on vast amounts of text data to process and generate human-like responses, answer follow-up questions, and complete other language-related tasks. There is widespread enthusiasm about the use of foundation models, especially LLMs and approaches that build on LLMs, to advance scientific research (Lee 2024).

When these models are used in scientific research, they encounter challenges including limited domain-specific knowledge, interpretability of the results, sparse training data, integration with experimental data, lack of causal understanding, and the evolving nature of scientific knowledge.

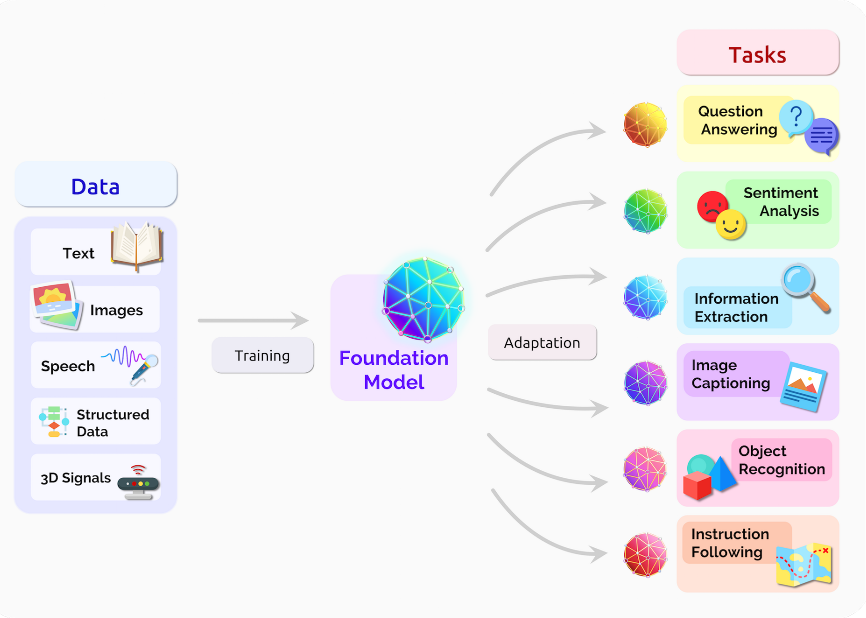

FIGURE 1-1 A foundation model centralizes the information from all the data from various modalities and can then be adapted to a wide range of downstream tasks. SOURCE: Bommasani et al. 2021. CC BY 4.0.

These challenges provide opportunities for research across all areas of DOE. The pursuit of these opportunities is an important endeavor as the private sector is presently leading the race for the development of state-of-the-art foundation models. The landscape of this race is in constant flux, and the leaders at any time will reap major rewards and may determine the direction of future scientific endeavors.

The DOE national laboratories are special-purpose entities referred to as federally funded research and development centers (FFRDCs). FFRDCs provide the government with a dedicated, objective, and highly specialized technical and analytical capability that is essential for addressing long-term, complex national challenges. FFRDCs cannot manufacture products or compete directly with industry and have no commercial or shareholder interests, ensuring that their advice, analysis, and research are unbiased, allowing them to act as "honest brokers" and trusted advisors. They attract, develop, and retain unique scientific expertise that combines world-class research and entrepreneurial know-how to support the mission of the agencies they serve. By assembling teams of experts from various fields, FFRDCs address multifaceted technical challenges that often require high-risk experiments and large facilities, such as supercomputers or light sources. FFRDCs play a crucial role in maintaining and advancing the nation's scientific and technical expertise in critical areas and facilitate technology transfer to the private sector. As such, the DOE national laboratories have an important role to play in advancing AI technologies, particularly AI foundation models for scientific discovery and innovation.

AI, particularly with the emergence of foundation models, is a transformative force poised to redefine future economies, national security, scientific discovery, global power dynamics, and daily life. Given this immense impact, maintaining U.S. leadership in AI is imperative, necessitating an understanding of the global competitive landscape, particularly coming from China.

China has strategically prioritized its development of AI, aiming to become a world leader in the field by 2030. This goal was outlined in its "Next-Generation Artificial Intelligence Development Plan," which was released in July 2017. Their ambition is supported by significant government investment in AI theory, technology, and application. Chinese AI firms have expanded their influence by freely distributing their models for the public to use, download, and modify, which makes them more accessible to researchers and developers around the world. In terms of quantifiable metrics, China is ahead of the United States: it significantly outpaced the United States in AI patent filings in 2022, possesses a leading advantage in the sheer volume of data, and leads the United States in the quantity of AI scientific papers. China has cultivated a robust domestic ecosystem, boasting abundant science, technology, engineering, and mathematics talent, resilient supply chains, and impressive manufacturing capabilities (Omaar 2024).

The nation that shapes the LLMs powering tomorrow's applications and services will wield great influence not only over the norms and values embedded in them but also over the critical semiconductor ecosystem that underpins AI computing. The fact that both China and the United States believe that these technologies could also provide military advantages only heightens the importance of achieving and maintaining long-term AI leadership.

Although the report will be examining use of foundation models for scientific discovery and innovation specifically for DOE, the development and use of these tools will benefit the general scientific community. The report will examine how foundation models can help drive progress in complex systems—such as digital twins—and unlocks new findings in areas vital to American competitiveness, including materials science, nuclear science, and public health.

STUDY APPROACH

The study was supported by DOE's Office of Science, National Nuclear Security Administration, and Biological and Environmental Research program. In collaboration with the National Academies of Sciences, Engineering, and Medicine, these DOE offices developed the study's statement of task (see Box 1-1). The National Academies appointed a committee of 11 members with expertise in mathematics, statistics, computer science, data science, algorithms and scal-

BOX 1-1
Statement of Task

A National Academies of Sciences, Engineering, and Medicine consensus study will assess the state of the art in foundation models and their use across science research domains relevant to the Department of Energy mission. The study will address the following questions:

- What are some exemplar use cases where foundation models could impact scientific discovery and innovation?
- How can foundation models be used in conjunction with traditional modeling, computational, and data science approaches?
- How can challenges such as verification, validation, uncertainty quantification, and reproducibility best be addressed to advance trustworthy foundation models?
- What are priority research areas for investments to advance the development and use of foundation models in scientific applications? What are the trade-offs in investing in foundation models versus other mathematical and computational approaches?

ability, energy consumption and computing, scientific applications, model trustworthiness, and DOE and laboratory experience. Committee biographies are provided in Appendix D.

The committee held several information-gathering meetings in support of this study, including one in-person public meeting (March 11–12, 2025) where the committee was presented with material from industry scientists and AI leaders from DOE laboratories. The other information-gathering sessions (February 11, May 6, and May 20, 2025) were virtual where presenters discussed DOE's interest in AI for science, learning models from data, and agentic AI.

Report Organization

This report was written with the intention of informing the scientific and research community, academia, pertinent government agencies, AI practitioners, and those in relevant industries about open needs when developing and using foundation models. The study takes an objective approach to understanding the field of foundation models specifically for scientific discovery and innovation and the potential opportunities that their use and development can bring to DOE. The report begins with a discussion on the use of foundation models with and without traditional modeling techniques[1] (Chapter 2). Chapter 3 explores the suc-

[1] For this report, traditional modeling refers to large-scale computational science solvers as well as statistical models.

cesses and exemplar use cases of foundation models and potential applications in which DOE could be most successful in its endeavors with foundation models for science. Chapter 4 discusses the strategic considerations and directions of foundation model use while challenges that the use of foundation models impose are covered in Chapter 5. The committee addresses major conclusions and recommendations throughout Chapters 2 through 5.

The committee would like to stress that while the report uses the terms AI, AI for science, AI models, LLMs, machine learning, and foundation models, the report is specifically directed toward the use and development of foundation models for science. The report is further specifically directed toward DOE's use and development of these models.

REFERENCES

Bommasani, R., D.A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M.S. Bernstein, et al. 2021. "On the Opportunities and Risks of Foundation Models." *arXiv*. https://doi.org/10.48550/arXiv.2108.07258.

Lee, A. 2024. "What Are Large Language Models Used For?" *NVIDIA Blog*. January 26. https://blogs.nvidia.com/blog/what-are-large-language-models-used-for.

Omaar, H. 2024. "How Innovative Is China in AI?" *Information Technology & Innovation Foundation*. August 26. https://itif.org/publications/2024/08/26/how-innovative-is-china-in-ai.

2

Foundation Models and Traditional Modeling

DEFINING THE SCOPE AND USE OF FOUNDATION MODELS

The landscape of artificial intelligence (AI) is undergoing a significant transformation driven by the emergence and evolution of foundation models. These models, typically large-scale neural networks trained on vast quantities of heterogeneous data, represent a departure from traditional AI systems designed for specific tasks. Foundation models possess the capacity to generate findings and discern patterns within extensive data sets with data volumes that exceed by orders of magnitude the computing and storage capacities of traditional solvers and even previous machine learning models.

Key characteristics defining foundation models include the following:

- *Massive scale:* Trained on vast data sets (web-scale, trillions-plus of data points) with immense internal complexity (trillions-plus of parameters), requiring significant computational resources for their processing.
- *Self-supervised pretraining:* Learning from unlabeled data, reducing the need for manual annotation.
- *Adaptability (transfer learning):* Easily fine-tuned for diverse downstream tasks, leveraging pretrained generalizable knowledge.
- *Emergent inference:* The ability to derive context and demonstrate reasoning that is not explicitly in the training data.
- *Multimodal and task-agnostic:* Ability to handle multiple modes of inputs, regardless of task.
- *Multipurpose architecture:* The architectures featuring combinations of transformers with attention mechanisms, encoders/decoders, and multilayer perceptrons are proving effective across modalities.

- *Scalability:* Performance generally improves with larger models, data sets, and computational capability. The same architecture can be adapted across domains via fine-tuning and deployed in resource-constrained environments using quantization (lower precision), selective activation of parameters, and low-rank adaptation.
- *Generalizability:* Transferring knowledge across diverse tasks and domains with minimal retraining, enabling strong zero-shot and few-shot performance, valuable for scientific and engineering applications using large technical data sets.

These characteristics position foundation models as a potential paradigm shift[1] for scientific research with a concomitant impact on the Department of Energy's (DOE's) mission.

BENEFITS OF ONLY USING FOUNDATION MODELS

In evaluating the roles of foundation models for scientific discovery, a natural early question is whether they present stand-alone alternatives to the modeling approaches that preceded them. We examine this perspective in the current section.

The key strengths of foundation models lie in adaptability, generalizability, scalability, and their capacity for multimodal integration. Foundation models can seamlessly combine multiple data modalities—including numerical simulation outputs, experimental sensor data, textual documentation, images, and videos—into unified representational frameworks. This unique capability makes them particularly well suited for fields such as life sciences, materials science, fluid dynamics, weather forecasting, and energy systems, where data complexity and heterogeneity pose significant challenges to traditional methods.

The scalability of foundation models, supported by large-scale computational resources, allows them to uncover complex patterns and interactions within massive data sets. This results in accelerated discovery and improved predictive performance in multifaceted scientific scenarios (Bodnar et al. 2025). Their generalized learning mechanisms further enable deployment across diverse operational contexts without requiring extensive manual reprogramming. In environments such as DOE facilities, this adaptability can lead to more dynamic and responsive control systems, enhancing operational efficiency and resilience in the face of evolving conditions.

Within DOE's computational science program, foundation models bring two particularly valuable advantages. The first centers on *spatiotemporal foundation models*—transformer-based architectures pretrained on large data sets derived

[1] A paradigm shift in this context means a fundamental change in how scientific research is conducted, driven by the introduction of foundation models.

from high-fidelity simulations of multiphysics systems described by partial differential equations. These models can forecast spatiotemporal solutions, aligning with one of the core goals of scientific machine learning: extending the capabilities of traditional, computationally expensive, discretization-based solvers. Spatiotemporal foundation models offer dramatic reductions in computational cost, enabling large-scale or long-time simulations at up to five orders of magnitude less computational effort. This has been compellingly demonstrated in domains such as Earth system modeling (Bodnar et al. 2025).

Although spatiotemporal foundation models may not yet achieve the accuracy of equation-based solvers, they offer a compelling trade-off through fine-tuning. When pretrained on a broad range of physics, these models can be adapted to new physical systems not present in the original training data. Remarkably, this transfer learning often yields better results than training the model from scratch on a single, narrow domain. Thus, spatiotemporal foundation models not only function as efficient solvers but also provide a scalable framework for generalizing across physical phenomena—an invaluable capability in a wide-ranging computational science program. Examples include multiphysics pretraining (McCabe et al. 2023) and co-domain neural operators (Rahman et al. 2024).

A second and perhaps even more intriguing potential lies in the inference of *emergent physics*. Because of the underlying transformer architecture—specifically the use of attention mechanisms and the capacity to learn contextual relationships over long pretraining epochs—these models may begin to reveal new physical findings or discoveries. They could go beyond simply generating solutions to explain the emergence of features in space and time, such as why vortex structures emerge in certain regions of a flow at specific times, or how macroscale material failure occurs as a consequence of microcrack and dislocation interactions. Such tasks are central to the roles of computational physicists. This possibility becomes even more plausible when spatiotemporal foundation models are integrated with large language models (LLMs) into multimodal systems (Ashman et al. 2024). Such combinations may bridge the gap between predictive modeling and interpretive reasoning, bringing us closer to models that not only solve complex physical systems but also explain them.

BENEFITS OF USING "TRADITIONAL MODELING" OVER FOUNDATION MODELS

In the context of DOE applications, *traditional models* refer to large-scale computational science solvers as well as statistical models. The solvers include finite element, finite difference, finite volume, and spectral methods and related numerical techniques. Over decades, partnerships between DOE and computational science researchers at U.S. universities have fostered the development of a robust ecosystem of discretization-based solvers. Supported by DOE's Advanced Scientific Computing Research and DOE's National Nuclear Security Administration Advance Simulation and Computing programs, this effort has produced vast suites of high-performance scientific software, much of it pioneered within DOE laboratories. Examples include the Trilinos Project, Dakota, and MFEM (Adams et al. 2020; Anderson et al. 2021; Heroux et al. 2005).

This ecosystem enables the modeling and simulation of a wide range of multiphysics problems relevant to DOE missions. It has evolved to support computations at the exascale and beyond, laying a firm foundation for applying computational science to complex, large-scale problems in physics, energy, Earth systems, and national security. As machine learning and artificial intelligence have grown in prominence, DOE-supported computational frameworks have begun to incorporate these data-driven methods, enriching traditional modeling approaches without discarding them.

A core strength of discretization-based solvers is their ability to deliver high-fidelity solutions that accurately represent the underlying physics—bounded mainly by the numerical algorithms and available computing power. These solvers explicitly encode conservation laws (e.g., energy, mass, momentum), thermodynamic consistency, and convergence properties, ensuring that model predictions are transparent, interpretable, and physically grounded. Such fidelity, however, comes at a cost: these models often demand significant computational resources, especially for large spatial domains or long-time horizons.

Despite the emergence of foundation models, traditional physics-based models retain critical advantages, particularly in interpretability, reliability, and strict adherence to physical laws. They are accompanied by rigorous verification, validation, and uncertainty quantification frameworks essential for DOE's high-stakes applications—such as nuclear reactor safety, weapons stewardship, and other national security tasks. These frameworks ensure compliance with safety, regulatory, and quality standards, which remain challenging for purely data-driven foundation models to satisfy. Furthermore, foundation models have yet to demonstrate generalizability across geometries, initial and boundary conditions and transitions such as phase changes, laminar-to-turbulent flow, shock formations, and material failure. These are standard for advanced discretization-based solvers.

In addition to being more amenable to interpretation and to the quantification of their uncertainty, traditional models often require less computational overhead for model development and deployment compared to the extensive pretraining and fine-tuning phases of foundation models. (However, geometry and mesh generation can prove time-consuming, and the expense of large direct numerical simulations is a well-recognized limitation.) Furthermore, traditional models play a foundational role in the data ecosystem—they are often required to *generate* the high-quality data used to train, fine-tune, or validate foundation models.

Another powerful advantage of traditional approaches lies in their ability to be integrated into statistical modeling frameworks. In many settings, physics-based models, of moderate or lower fidelity, can be embedded within Bayesian hierarchical structures to facilitate efficient uncertainty quantification.

BENEFITS OF USING TRADITIONAL MODELING WITH FOUNDATION MODELS

Integrating Foundation Models with Traditional Scientific Computing: A Pathway to Accelerated Discovery

Integrating traditional modeling approaches with foundation models offers transformative potential for DOE's scientific enterprise. Traditional computational methods—such as finite element, finite volume, and spectral solvers—have formed the bedrock of high-fidelity simulations, enabling predictive science across complex domains such as materials physics, turbulent fluid flow, Earth systems modeling, and nuclear systems, as outlined above. These models are grounded in well-understood physical laws and verification and/or validation protocols, making them indispensable for safety-critical and regulatory-constrained applications. However, they come with significant computational demands, particularly for large-scale or long-time simulations.

The committee reiterates that by contrast, foundation models trained on vast multimodal data sets—including simulation results, sensor data, imagery, and scientific literature—offer scalability, generalizability, and data-driven adaptability. Rather than viewing foundation models as replacements for traditional methods in computational science, the committee advocates for a *synergistic integration* of the two (Koumoutsakos 2024). Hybrid modeling strategies can fuse the physical interpretability and robustness of classical solvers with the efficiency and learning capabilities of foundation models, particularly in multiscale, multiphysics applications where stand-alone approaches often fall short.

Accelerating Scientific Discovery Through Hybrid Approaches

Foundation models can significantly enhance the entire research life cycle at DOE national laboratories and user facilities through multiple avenues of hybridization:

- *Simulation Acceleration and Enhancement:* Foundation models, trained as surrogate models, can emulate computationally expensive physics simulations, allowing for accelerated parameter sweeps, ensemble studies, and real-time forecasting. Applications range from turbulence and fusion modeling to Earth systems science and high-energy physics. Moreover, foundation models can discover governing dynamics—such as learning coefficients in ordinary differential equations or structures in partial differential equations (PDEs)—directly from data (Ye et al. 2025). This enables breakthroughs in both forward prediction and inverse problem-solving (Bodnar et al. 2025; McCabe et al. 2024; Nguyen et al. 2023; Rahman et al. 2024; Ye et al. 2025).

- *Experimental Data Analysis:* DOE facilities generate massive data sets across diverse modalities. Multimodal foundation models can interpret these data in real time, performing automated feature extraction, anomaly detection, and pattern recognition—for example, identifying material phases in scattering data (Brodnik et al. 2023). This capability paves the way for "self-driving" experiments that optimize limited facility time and dynamically adjust to emergent results, fundamentally transforming experimental workflows.
- *Knowledge Discovery and Hypothesis Generation:* With the scientific literature growing exponentially, foundation models—especially LLMs fine-tuned on curated corpora such as DOE's Office of Scientific and Technical Information repositories (Sakana.AI, 2024; Skarlinski et al. 2025)—can synthesize findings, identify knowledge gaps, generate novel hypotheses, and suggest experiment designs. Programs such as the Defense Advanced Research Projects Agency's Discovery of Algorithms and Architectures illustrate how LLMs can discover fundamental scientific computing modules, further validating the utility of AI in hypothesis-driven research (DARPA 2025).
- *Autonomous Laboratories:* The fusion of foundation models with robotics and automated platforms unlocks the vision of "self-driving laboratories" that can autonomously design, execute, and interpret experiments (Skarlinski et al. 2025). These systems promise to dramatically accelerate research cycles in materials discovery, synthetic biology, and beyond.

Methods of Integration: Hybrid and Agentic Architectures

Foundation model development is progressing toward augmenting traditional simulations by learning data-driven corrections to reduced-order models. For example, foundation model–based closure approximations in turbulence and combustion science could improve fidelity, while in nuclear and Earth systems modeling, they could enhance accuracy and enable rigorous uncertainty quantification:

- *Data-Driven Corrections:* Foundation models can augment traditional simulations by learning data-driven corrections to approximate or simplified models. For example, foundation model–based closure approximations in turbulence and combustion science improve fidelity, while in nuclear and Earth systems modeling, they enhance accuracy and enable rigorous uncertainty quantification (Bodnar et al. 2025).

- *Inference from Traditional Simulations:* A direct method of integration involves generating spatiotemporal simulations using traditional solvers and then using LLMs for inference. Given snapshots of numerical fields—as either gridded data or imagery—current-generation LLMs can describe and interpret system behaviors in natural language. This capability is further enhanced by combining simulation outputs with symbolic and textual representations of the underlying physics (McCabe et al. 2023; Rahman et al. 2024).
- *Spatiotemporal Foundation Models:* These models, pretrained and fine-tuned on outputs from traditional solvers, learn spatiotemporal variation from their training data while enabling rapid forecasts in new contexts. Their ability to generalize across physics, especially when pretrained on diverse PDEs and fine-tuned to specific ones, highlights the value of transfer learning in computational science (Herde et al. 2024; McCabe et al. 2023; Rahman et al. 2024).
- *Agentic Workflows with Reasoning Capabilities:* Multimodal LLMs can orchestrate workflows where AI agents dynamically choose between invoking a traditional solver or using a pretrained foundation model. These agents integrate simulations, mathematical formulations, and natural language descriptions to perform inference, design studies, or explain observed behaviors. Advanced techniques such as retrieval-augmented generation and reasoning further improve performance by grounding reasoning in contextually relevant information (Gottweis et al. 2025).

Toward a New Scientific Paradigm

The fusion of foundation models with traditional numerical methods represents more than a computational advance: it constitutes a paradigm shift in how scientific discovery is conducted. By combining rigorous physical modeling with the adaptive learning capabilities of modern AI, this hybrid approach opens the door to *faster, more accurate, and more autonomous science*.

From accelerating simulations to enabling real-time experimental feedback and automating hypothesis generation, the integration of foundation models into DOE's computational and experimental ecosystem promises to reshape the pace and scope of scientific innovation (Bodnar et al. 2025; Herde et al. 2024; McCabe et al. 2023; Nguyen et al. 2023; Ye et al. 2024, 2025).

> *Conclusion 2-1: Integrating traditional models with foundation models is proving to be increasingly powerful and has significant potential to advance computational findings in the physical sciences. These hybrid methods leverage the physical interpretability and structures of classical computational approaches alongside the data-driven adaptability of foundation models. This integration enables the modeling of*

complex multiphysics, multiscale, and partially observed (understood) systems that challenge traditional approaches both computationally and mathematically.

Recommendation 2-1: The Department of Energy (DOE) should invest in foundation model development, particularly in areas of strategic importance to DOE, including areas where DOE already has advantages leveraging its unique strengths in those domains. DOE should also prioritize the hybridization of foundation models and traditional modeling. Such hybrid modeling strategies can fuse the physical interpretability and robustness of classical solvers with the efficiency and learning capabilities of foundation models, particularly in multiscale, multiphysics applications where traditional approaches have limitations in capturing the heterogeneity, complexity, and dynamics of the physical system. DOE should not, however, abandon its expertise in numerical and computational methods and should continue investing strategically in software and infrastructure.

REFERENCES

Adams, B.M., W.J. Bohnhoff, K.R. Dalbey, M.S. Ebeida, J.P. Eddy, M.S. Eldred, R.W. Hooper, et al. 2020. *Dakota, a Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 6.13 User's Manual*. Sandia National Laboratories. https://www.sandia.gov/app/uploads/sites/241/2023/03/Users-6.13.0.pdf.

Anderson, R., J. Andrej, A. Barker, J. Bramwell, J.S. Camier, J. Cerveny, V. Dobrev, et al. 2021. "MFEM: A Modular Finite Element Methods Library." *Computers and Mathematics with Applications* 81:42–74.

Ashman, M., C. Diaconu, E. Langezaal, A. Weller, and R.E. Turner. 2024. "Gridded Transformer Neural Processes for Large Unstructured Spatio-Temporal Data." *arXiv*:2410.06731, https://ui.adsabs.harvard.edu/abs/2024arXiv241006731A, accessed October 1, 2024.

Bodnar, C., W.P. Bruinsma, A. Lucic, M. Stanley, A. Allen, J. Brandstetter, P. Garvan, M. Riechert, J.A. Weyn, H. Dong, J.K. Gupta, K. Thambiratnam, A.T. Archibald, C.C. Wu, E. Heider, M. Welling, R.E. Turner, and P. Perdikaris. 2025. "A Foundation Model for the Earth System." *Nature* 641(8065):1180–1187.

Brodnik, N.R., C. Muir, N. Tulshibagwale, J. Rossin, M.P. Echlin, C.M. Hamel, S.L.B. Kramer, T.M. Pollock, J.D. Kiser, C. Smith, and S.H. Daly. 2023. "Perspective: Machine Learning in Experimental Solid Mechanics." *Journal of the Mechanics and Physics of Solids* 173:105231. https://doi.org/10.1016/j.jmps.2023.105231.

DARPA (Defense Advanced Research Projects Agency). 2025. "DIAL: Mathematics for the Discovery of Algorithms and Architectures." https://www.darpa.mil/research/programs/mathematics-for-the-discovery-of-algorithms-and-architectures, accessed July 31, 2025.

Gottweis, J., W.-H. Weng, A. Daryin, T. Tu, A. Palepu, P. Sirkovic, A. Myaskovsky, et al. 2025. "Towards an AI Co-Scientist." *arXiv*:2502.18864 (eprint). https://ui.adsabs.harvard.edu/abs/2025arXiv250218864G.

Herde, M., B. Raonić, T. Rohner, R. Käppeli, R. Molinaro, E. de Bézenac, and S. Mishra. 2024. "Poseidon: Efficient Foundation Models for PDEs." *arXiv*:2405.19101. https://ui.adsabs.harvard.edu/abs/2024arXiv240519101H.

Heroux, M.A., R.A. Bartlett, V.E. Howle, R.J. Hoekstra, J.J. Hu, T.G. Kolda, R.B. Lehoucq, et al. 2005. "An Overview of the Trilinos Project." *ACM Transactions on Mathematical Software* 31(3):397–423.

Koumoutsakos, P. 2024. "On Roads Less Travelled Between AI and Computational Science." *Nature Reviews Physics* 6(6):342–344.

McCabe, M., B. Régaldo-Saint Blancard, L. Holden Parker, R. Ohana, M. Cranmer, A. Bietti, M. Eickenberg, et al. 2023. "Multiple Physics Pretraining for Physical Surrogate Models." *arXiv*:2310.02994. https://ui.adsabs.harvard.edu/abs/2023arXiv231002994M (last revised December 10, 2024).

Nguyen, T., J. Brandstetter, A. Kapoor, J.K. Gupta, and A. Grover. 2023. "ClimaX: A Foundation Model for Weather and Climate." *arXiv*:2301.10343. https://ui.adsabs.harvard.edu/abs/2023arXiv230110343N.

Rahman, M.A., R.J. George, M. Elleithy, D. Leibovici, Z. Li, B. Bonev, C. White, et al. 2024. "Pretraining Codomain Attention Neural Operators for Solving Multiphysics PDEs." *arXiv*:2403.12553. https://doi.org/10.48550/arXiv.2403.12553.

Sakana.AI. 2024. "The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery." https://sakana.ai/ai-scientist.

Skarlinski, M., T. Nadolski, J. Braza, R. Storni, M. Caldas, L. Mitchener, M. Hinks, A. White, and S. Rodriques. 2025. "FutureHouse Platform: Superintelligent AI Agents for Scientific Discovery." FutureHouse. https://www.futurehouse.org/research-announcements/launching-futurehouse-platform-ai-agents.

Ye, Z., X. Huang, L. Chen, H. Liu, Z. Wang, and B. Dong. 2024. "PDEformer: Towards a Foundation Model for One-Dimensional Partial Differential Equations." *arXiv*:2402.12652. https://ui.adsabs.harvard.edu/abs/2024arXiv240212652Y.

Ye, Z., Z. Liu, B. Wu, H. Jiang, L. Chen, M. Zhang, X. Huang, et al. 2025. "PDEformer-2: A Versatile Foundation Model for Two-Dimensional Partial Differential Equations." *arXiv*:2507.15409. https://ui.adsabs.harvard.edu/abs/2025arXiv250715409Y.

3

Exemplar Use Cases of Foundation Models

DEPARTMENT OF ENERGY'S ROLE IN FOUNDATION MODEL DEVELOPMENT

The strategic focus of a Department of Energy (DOE)-wide foundation model initiative remains a subject of debate, requiring the department to balance its broad application space, navigate the trade-offs between leveraging past industry advancements, address the unique national security imperatives of DOE, and ensure responsible stewardship of taxpayer resources, particularly in light of the opportunity costs associated with prioritizing artificial intelligence (AI) over more mature technologies.

There are an ever-increasing number of efforts across DOE national laboratories integrating AI and foundation models into their research programs. Naturally, one of the key targets is energy-related applications ranging from electric grids to nuclear fusion. A primary consideration is the perception that DOE cannot compete with the head start in technology maturation and large market share currently held by large companies. In the foundation model market, leaders such as Microsoft (via OpenAI), Google/Gemini, Amazon Web Services, Meta, and Anthropic each back efforts with investments ranging from $10 billion to more than $75 billion in funding and infrastructure (Fernandez et al. 2025), a scale that DOE would be hard pressed to match. This raises a fundamental question of whether DOE should focus on collaborations with industry or focus on a complementary space based on DOE's unique mission in curating foundational science while improving national security. The committee believes that DOE has reason to develop foundation models internally, *in addition* to private-sector leadership, because the needs of the government (whether for national security or continued scientific preeminence) will not be met by private interests. The two endeavors (private and public) do not compete—they complement each other and can leverage each other.

> *Conclusion 3-1: Commercial industry has driven rapid progress in developing large language model–based foundation models, yielding a robust ecosystem of tools and capabilities. As demonstrated by, for example, the collaboration between Los Alamos National Laboratory and OpenAI, DOE can leverage industry advances, findings, and collaborations as it develops foundation models for science and conducts coordinated DOE-wide assessments to identify appropriate opportunities.*

DOE is the largest single federal sponsor of scientific research in the United States, providing approximately $16 billion in research and development (R&D) funding in fiscal year (FY) 2023, which represents roughly 8 percent of total federal R&D obligations (Blevins 2022). Through its Office of Science, DOE supports approximately 40 percent of all federal basic research in the physical sciences, and an estimated 44 percent of federal basic research in computer and information sciences, including foundational work in nonconvex optimization, probabilistic methods, and large-scale high-performance computing (NCSES 2023). Although DOE is unlikely to match the pace or scale of commercial product development, it retains clear strategic advantages in five areas: (1) a world-class scientific workforce in computational science; (2) access to large-scale, science-focused, and experimental computing hardware; (3) stewardship of unique experimental facilities and open and controlled or classified scientific data; (4) capability to tackle long-term, high-risk, high-reward scientific problems; and (5) access to unique scientific data that may not be easily reproduced and which can be expanded as synthetic data may be necessary for training future foundation models.

Despite a mismatch in funding allocations, DOE's Exascale Computing Project (ECP)[1] guided the development, procurement, and construction of the Frontier and Aurora supercomputers at a total cost of approximately $1.7 billion. Frontier achieves a peak performance of 1.35 exaflops, and Aurora reaches approximately 1.01 exaflops. For a rough comparison, a machine like Aurora could train a model like GPT-4 on the order of ~200 days, suggesting that the best option for DOE is not to directly compete in the same general-purpose paradigm. With recent attention toward the disruption of DeepSeek, which some analyses suggest offered a ~10× increase in efficiency, existing ECP-funded resources become arguably more competitive, particularly when buoyed by the highly skilled workforce represented by the national laboratories. In fact, ECP-funded resources have the potential to train foundation models from scratch, deploy stochastic optimization algorithms at scale, or run multiagent simulations in real time. This is evidence that the field is advancing in a direction that could make DOE's resources feasible for the training of foundation models for science.

[1] See https://www.exascaleproject.org.

When comparing DOE and industry capabilities, one of the most significant differences is the scale and richness of physical and simulation data generated by DOE's network of user facilities, nuclear weapons–testing archives, ongoing experimental campaigns, and high-performance computing facilities. These data span both classified and unclassified domains and present unique opportunities for DOE-relevant advances in foundation models. New modes of inquiry that have become successful in the industrial setting (e.g., Google DeepMind's alpha-evolve) have great potential for DOE applications. A central technical challenge is whether secure training methods, such as federated learning, can be designed to mathematically preclude the leakage of sensitive or controlled information. If so, this could enable the construction of scientific foundation models that operate across heterogeneous and compartmentalized data sources. If not, certain classes of model architectures may prove fundamentally incompatible with DOE's mission constraints.

> *Conclusion 3-2: DOE retains clear strategic advantages in five areas: (1) a world-class scientific workforce in computational science; (2) access to large-scale, science-focused, and experimental computing hardware; (3) stewardship of unique experimental facilities and open and controlled or classified scientific data; (4) capability to tackle long-term, high-risk, high-reward scientific problems; and (5) access to unique scientific data that may not be easily reproduced and which can be expanded as synthetic data may be necessary for training future foundation models.*

Many of DOE's experimental platforms are already compatible with remote operation and automation. This includes user-facing beamlines, additive manufacturing facilities, and autonomous platforms for chemical synthesis and materials fabrication. At the same time, recent advances in retrieval-augmented generation (RAG) have introduced new strategies for connecting large language model (LLM) outputs with authoritative external sources. DOE could consider a coordinated program, either independently or in partnership with academia and industry, in which traditional physics-based simulations or experiments are launched in an agentic loop and used to refine LLM reasoning. This concept is particularly viable in diverse domains such as small-molecule chemistry and mature simulation codes (e.g., computational fluid dynamics, electromagnetism, molecular dynamics). An important example is the recent effort by researchers at Lawrence Livermore National Laboratory to combine AI with fusion target design by deploying AI agents on two of the world's most powerful supercomputers to automate inertial confinement fusion simulations and thus accelerate experiments. Additional potential benefits of AI in the quest for fusion energy are provided in the next section.

However, in many scientific contexts, human expertise remains essential for initiating, interpreting, and validating results. Discovery via the use of experimental or computational platforms relies crucially on the deep bench of technical expertise at the laboratories that can be rapidly tapped to analyze previously unseen scenarios in high-consequence national security settings with limited time to solution. In these settings, foundation models may act as an accelerant for analysis, but are currently not viewed as sufficiently reliable for trustworthy application.

HUMAN IN THE LOOP AND ARTIFICIAL INTELLIGENCE FOUNDATION MODEL AUTONOMY

Human oversight remains essential in deploying and utilizing foundation models, especially in high-risk or high-impact scientific and engineering contexts. Foundation models enhance the productivity of researchers by, for example, accelerating targeted literature reviews, optimizing code and algorithm design, and dramatically reducing the time required to prototype and validate solutions. By automating many of the routine or well-established steps in the research process, foundation models allow scientists and engineers to focus on higher-level reasoning and innovation. However, it is important to keep in mind that these capabilities come with a significant caveat: foundation models are capable of generating both highly sophisticated statements and nonsense. Although they can produce novel findings and accurate solutions, they can just as easily generate plausible-sounding but incorrect or misleading outputs. For this reason, and for the foreseeable future, a human-in-the-loop approach is desirable (even essential), ensuring that domain expertise and critical thinking guide the use and interpretation of model outputs. In this context, we mention that there are different levels or schemas of handoff, that is, the transfer of decision-making authority or control between a human and a foundation model (or agentic environment). The nature of the handoff depends on both the confidence in the model's output as well as on the level of criticality (risk) associated with the decision one is trying to make. Importantly, this concept and associated schemas will evolve as foundation models become more mature and trustable (see Chapter 5 for details on quantifiable confidence).

A powerful example of this human–AI interaction is DeepMind's AI co-scientist work (Gottweis and Natarajan 2025), a multiagent system built on Gemini, designed to assist scientists, engineers, and researchers in general in formulating hypotheses, conducting literature reviews, and building experimental frameworks. In this work, specialized agents operate asynchronously to generate, evaluate, and fine-tune scientific hypotheses. In several instances, this collaborative approach has made it possible for scientists to interact easily and very naturally with AI, providing inputs, prompts, or feedback to guide research, with final oversight and authority remaining with the investigator. For example, the AI Co-Scientist has demonstrated its potential impact in biomedical research, suggesting novel approaches to inhibit disease progression in conditions, such as liver fibrosis, that showed promising potential.

Another compelling example of human–AI interaction enabled by foundation models is the use of AI copilots in software development. Importantly, this approach is quickly becoming the norm in modern software engineering. For example, tools such as Cursor, which is built on top of LLMs and fine-tuned for code generation, offer real-time support in writing, debugging, and refactoring code (Anysphere n.d.). These systems serve as smart and efficient collaborators, helping developers implement complex algorithms and explore alternative code-design patterns. Some of these tools integrate seamlessly with developer workflows, allowing users to query codebases in natural language, generate multifile implementations, and suggest algorithmic solutions from minimal user prompts and/or examples. Although these tools do not replace developers, they act as accelerators, cutting down on repetitive and established coding tasks. As a result, engineers can focus on architectural decisions and problem solving. Notably, this also lowers software skill requirements. Again, the human remains in the loop; although the model may generate functional code, oversight is mandatory to validate correctness.

This paradigm illustrates how the synergy between human expertise and foundation model capabilities can lead to more efficient, reliable, and responsible scientific outcomes.

> *Conclusion 3-3: While AI systems can exceed human performance in many ways, they can also fail in ways a human likely never would. For this reason, the qualification of AI will be necessary for decision making and prediction in the presence of uncertainty.*

Based on the discussion above, the integration of foundation models into scientific and engineering pipelines raises concerns about the future of employees working in these sectors. In fact, although these models boost productivity, they put at risk those roles that are focused on repetitive, manual, or routine tasks. Roles dedicated to basic literature reviews, boilerplate coding, standard documentation, and straightforward data analysis could be significantly affected. Importantly, in many cases, foundation models could be able to perform at scale and with more reliability than that of a human.

On the other hand, roles that require deep domain expertise and critical judgment (e.g., principal investigators, senior engineers, code architects, and regulatory or quality assurance engineers) are less likely to be removed. In fact, because of the need for human oversight when it comes to the interpretation of output of foundation models, these roles become even more valuable, as they are fundamental in verifying and building on top of what AI-based machines can achieve.

In short, foundation models potentially introduce a paradigm shift, where humans act as big-picture strategists and critical evaluators of AI-generated outputs, ensuring that they are technically correct and aligned with larger scientific and engineering goals.

Recommendation 3-1: The Department of Energy (DOE) should study and develop the fusion of artificial intelligence (AI) and human capabilities. At present, AI systems handle the repetitive, manual, or routine tasks, and are starting to show abilities to reason. As AI becomes more capable, deep analysis and strategy recommendations become feasible, but humans should maintain oversight and validation, particularly for qualification and other aspects of DOE's mission.

Recommendation 3-2: The Department of Energy should evaluate the capabilities and risks of agentic artificial intelligence (AI) systems for its core applications. In particular, the committee advocates exploring agentic AI for developing autonomous laboratories for scientific discovery, decision making, and action planning for high-stakes applications.

SCIENTIFIC AND ENGINEERING APPLICATIONS

In the context of scientific and engineering applications, foundation models (FMs) trained on observations, scientific literature, databases, as well as experimental results and outputs of simulations can be used to support hypothesis generation. In engineering, the use of FMs is becoming predominant in design settings, specifically in tasks such as CAD (computer-aided design) generation. These applications demonstrate how FMs can serve as intelligent copilots for researchers and engineers, enhancing productivity and enabling new modes of discovery.

DOE's mission encompasses many areas including materials science, chemistry, physics, energy, Earth systems, and high-performance computing, to name a few. DOE also supports national security missions such as stewardship of the nation's nuclear stockpile. Because DOE's mission includes so many scientific and engineering disciplines, it is only possible to provide a few examples below to illustrate how FMs might accelerate progress.

Materials Science

Materials science seeks to understand and control the relationships between structure, processing, properties, and performance across multiple spatiotemporal scales. FMs trained on experimental data, literature, and simulations offer a promising path to accelerate discovery—namely, through property prediction, retrosynthesis, and molecular generation. These models can predict properties, generate candidate structures, and guide automated experiments, reducing reliance on costly first-principles calculations (Berger 2025; Pyzer-Knapp et al. 2025). When coupled with high-throughput synthesis, they could transform materials discovery from a decades-long process into an iterative, data-driven cycle. Recent advances include MatBERT, a materials science transformer developed at Lawrence Berkeley National Laboratory (Trewartha et al. 2022), and IBM's open-source FMs for sustainable materials design (Martineau 2024). We highlight some areas in which FMs are already being used to significant impact—namely, property prediction, retrosynthesis, and molecular generation—and also look to the future to outline areas that we believe are key to continuing to unlock value. These areas hinge on exploiting the natural multimodality and multifidelity characteristics of materials data through increasingly powerful and elegant modeling approaches. The application of FMs faces challenges such as vast chemical and structural design spaces, bridging scales, and integrating experimental, computational, and theoretical insights into predictive frameworks with quantified uncertainties (Morgan and Jacobs 2020).

Battery Technology

In battery technology, FMs are accelerating innovation from materials to management. Researchers are developing these models to rapidly screen and predict the properties of novel battery materials, such as new electrolytes, thus speeding up the discovery process (Xu et al. 2024). Furthermore, they are being used to create more sophisticated battery management systems that provide highly accurate predictions of a battery's state of health and remaining useful life (Chan et al. 2025). By understanding the deep patterns of battery degradation, these models are helping to design safer, longer-lasting, and more efficient energy storage solutions for everything from electric vehicles to grid-scale applications.

Advanced Manufacturing

Advanced manufacturing (AM) uses computer-controlled, automated processes to produce complex components relevant to DOE's mission. This type of manufacturing distinguishes itself from conventional mold-based or subtractive manufacturing in that it enables rapid prototyping, cost-effective experimentation, and just-in-time production of complex components as a single unit (e.g., rocket nozzles). AM is increasingly vital to DOE and its National Nuclear Security Administration (NNSA) for both energy science and national security missions, with the goal of creating parts that are "born qualified" for their intended use (Boyce 2016). Moreover, FMs offer promising solutions by integrating heterogeneous data to support tasks such as anomaly detection, process optimization, and predictive control (Autodesk 2025; Era et al. 2025; NVIDIA n.d.; Zhang et al. 2025).

Key challenges remain, as AM materials are often out of thermodynamic equilibrium, leading to undesirable properties such as low ductility or fracture toughness (Forien 2023). Developing digital twins, computational replicas of AM processes, is a major research focus across DOE and NNSA laboratories (LLNL n.d.) and an area that is being transformed by the adoption of FMs.

Weather and Earth Systems

Predicting weather and understanding Earth systems is critical to decision making (Conti 2024). Large-scale simulations of weather remain limited in the range of their time horizons and spatial resolutions due to computational constraints. FMs offer solutions by developing accurate parameterizations for subgrid-scale processes such as clouds and turbulence, identifying patterns, and improving understanding of climate dynamics. FMs for Earth systems and weather are pretrained on massive, heterogeneous Earth-system data sets and are fine-tuned for diverse downstream tasks. Data-driven weather models, such as GraphCast (Lam et al. 2023) are becoming central to several FMs that involve localized forecasts and can significantly impact applications such as agriculture and power grids. Aurora (Bodnar et al. 2025), a 1.3-billion-parameter model pretrained on more than 1 million hours of multimodal geophysical data, outperforms traditional numerical forecasts in global weather forecasting, air quality monitoring, ocean wave prediction, and tropical cyclone tracking, all at lower computational cost. Similarly, Prithvi WxC (Schmude et al. 2024), a 2.3 billion-parameter transformer model trained on 160 atmospheric variables from MERRA-2, is designed for multitask adaptation, including downscaling, extreme-event estimation, and parameterization. Projects such as ORBIT (Wang et al. 2024), a hybrid transformer model with 113 billion parameters for Earth system predictability, hold potential to accelerate climate projections, improve extreme-event forecasts, and unify disparate Earth-system modeling tasks. Key challenges include integrating real-time data assimilation, maintaining physical consistency over long prediction horizons, and scaling to capture multiscale interactions across atmosphere, ocean, land, and cryosphere.

Fusion

The quest for fusion energy involves extremely complex plasma physics and engineering challenges. FMs can accelerate the computationally demanding simulations of plasma behavior (e.g., using codes such as X-Point included Gyrokinetic Code) (Churchill 2024), help analyze the vast amounts of diagnostic data from experiments such as Doublet III D-Shaped or the International Thermonuclear Experimental Reactor (see also the agenda for the Simple Cloud-Resolving E3SM Atmosphere Model), assist in designing reactor components tolerant of extreme conditions, and potentially contribute to real-time plasma control systems needed for sustained fusion reactions. We believe that this trend will continue. In inertial confinement fusion, AI and fine-tuned FMs can help design reproducible high-fusion-gain targets. These models are pretrained on vast and diverse data sets including experimental data from tokamaks and massive simulations. They are fine-tuned for downstream tasks, such as predicting plasma disruptions, optimizing control systems in real time, improving diagnostic interpretation, and accelerating the design cycle for reactor components

(Badalassi 2023; Churchill 2024; DOE 2024). In magnetic confinement experiments, such models are being explored for real-time control to adjust magnetic fields and mitigate instabilities like tearing modes, which can severely limit performance (DOE n.d.). Beyond control, LLMs augmented with (RAG are being used to quickly access historical data and identify similar experimental conditions to guide new trials (Poore 2023). Furthermore, the design of durable fusion materials and tritium-breeding blankets that can withstand extreme reactor environments is being addressed by integrating foundation models with high-performance computing to create comprehensive simulation environments (Badalassi et al. 2023; DOE 2024; PNNL n.d.).

The committee would like to stress some of the dangers of adapting an application too quickly. In the past 2 years, several groups have sought to replace costly plasma simulations with autoregressive neural surrogates that evolve hydrodynamic and electromagnetic fields without direct partial differential equation solutions (Carey et al. 2024, 2025; Galletti et al. 2025; Gopakumar et al. 2023; Poels et al. 2023). While these are important first steps, there are fundamental challenges that must be addressed before such approaches can form the basis of a true FM for fusion, comparable to those emerging in weather forecasting. Current efforts rely heavily on Fourier neural operators (FNOs), which cannot readily accommodate the complex geometries required for magnetic confinement fusion. Moreover, autoregressive roll-outs are prone to compounding errors over long prediction horizons (McCabe et al. 2023). This issue is particularly acute in fusion, where predictions must preserve gauge symmetries and conservation laws; this is well known in conventional plasma simulation contexts within the DOE community (Sharma et al. 2020). Off-the-shelf FNOs and transformer models lack these structural guarantees. A viable FM for fusion will therefore require new approaches that ensure long-term stability and strict preservation of physical structure.

Stockpile Stewardship

The U.S. Stockpile Stewardship Program (SSP), managed by NNSA and its nuclear enterprise, aims to maintain the safety, security, and reliability of the nuclear arsenal without resuming underground testing. The national laboratories involved in these efforts have made significant progress using machine learning to obtain a deeper understanding of the relevant science and are increasingly exploring the use of FMs. These FMs are tuned with classified weapons science knowledge to gain a deeper understanding of the physics involved, thereby accelerating progress across the entire program. This represents a substantial shift toward data-driven maintenance of the stockpile.

One critical area for FM deployment is stockpile surveillance, the continuous monitoring of the health of the arsenal. FMs can be fine-tuned using a wealth of past findings and diagnostic images to rapidly assess potential deleterious changes, helping experts quickly distinguish between changes that are not material to future performance and those that require intensive investigation via simulation and experiment. Furthermore, FMs are essential in designing digital twins that predict component failure over time—an especially difficult task. By measuring a part's response to its dynamic environment and assimilating these data, an FM can construct a digital twin to provide advance warnings of impending failure, such as fracture due to fatigue, allowing for proactive maintenance. There is significant overlap with areas such as structural health monitoring that may be useful to adopt in this effort (see following paragraph).

Despite their potential, the use of FMs for stockpile stewardship involves significant risks and challenges. The most prominent concern is security, as classified information must be strictly controlled and only provided to staff with the necessary "need to know" clearance, a protocol that must be maintained even within secure laboratory confines. Another challenge is preventing overreliance on the guidance provided by FMs, as this could inadvertently lead to poor design decisions regarding weapon components. The DOE laboratories involved in the SSP are well aware of these issues and are actively working to mitigate these risks.

Structural Health Monitoring

FMs are gaining significant attention for structural health monitoring and infrastructure surveillance, extending their utility from high-security areas such as the nuclear SSP to civilian applications such as bridges, viaducts, and high-rise buildings. FMs can absorb massive, unlabeled data sets derived from sensors—including accelerometers for vibration monitoring, imaging diagnostics, and Internet of Things devices. This generalized pretraining allows the models to learn robust, universal representations of structural behavior. Downstream tasks include anomaly detection and traffic load estimation on real-world civil infrastructure data (Benfenati et al. 2025; Bormon 2025; Hassani et al. 2024). A key application of FMs in civil infrastructure is the creation of intelligent, high-fidelity digital twins. By continuously assimilating real-time data from the physical structure (the "real twin"), FMs enable the virtual replica to accurately predict degradation, fatigue, and component failure over time. The integration of FMs into digital twins is an active area of investigation, aiming to reduce the significant manual effort typically required to create and maintain these models for cyber–physical systems (Ali et al. 2024). Although this technology promises enhanced safety and optimized resource allocation by distinguishing critical changes from nonmaterial ones, the field faces challenges related to data security, ensuring the fidelity and trustworthiness of FM-generated predictions, and managing the large computational resources required for both training and real-time inference.

Combustion

Combustion systems, including engines, gas turbines, furnaces, and scramjets, show highly unsteady, multiscale dynamics. These dynamics stem from complex interactions among turbulence, multiphase, and reacting flows. Current physics-based simulations are too costly for extensive design or operating space exploration and cannot directly use real-world experimental data. FMs are increasingly adopted for combustion research by leveraging vast heterogeneous data sets, such as direct numerical simulations, large-eddy simulations, and experimental diagnostics, to learn universal representations of combustion phenomena (Ihme and Chung 2024). FMs can assist in the acquisition of new insights into the physics controlling flame ignition, burning rate, flame stability, and emissions in high-pressure premixed combustion of various fuels, including hydrogen. These developments are crucial for the improvement of multifidelity science-based reduced-order models, methods, and digitalization, ultimately used by U.S. industry and its clients for optimal design and operation, near-real-time risk mitigation, and maintenance. Examples of ongoing efforts include a knowledge processing framework for combustion science that integrates FMs with RAG to systematically parse literature, data sets, and simulation results, enabling automated reasoning and accelerated model development (Sharma and Raman 2024). The interfacing of combustion and machine learning is mostly focused now on adopting supervised and semi-supervised machine learning techniques to combustion problems,

Recent progress in physics-informed machine learning provides a pathway to embedding physical constraints directly into FMs, making them suitable for high-fidelity combustion simulations (Cao et al. 2026). The adoption of an inverse modeling approach (Karnakov et al. 2024) and the extension of these efforts in order to account for proper validation and verification (McGreivy and Hakim 2024) within an FM framework holds great potential for combustion science, an area central to the mission of DOE.

National Security

In addition to the potential benefits described above, FMs can bolster other national security missions where DOE plays an important role:

- **Nonproliferation and threat detection.** FMs can process large, heterogeneous data sets (e.g., satellite imagery, sensor data) to identify nuclear proliferation activities or emerging threats.
- **Strategic analysis.** They can assist analysts by synthesizing information from technical, geopolitical, and open-source materials to support strategic decision making.

FMs offer powerful tools for managing and securing energy infrastructure, such as the following:

- **Grid management and optimization.** FMs trained on operational data, weather patterns, and energy markets can enhance load forecasting, predict renewable generation (solar, wind), and optimize grid operations for efficiency and stability.
- **Resilience and threat mitigation.** By analyzing complex system interdependencies, FMs can identify vulnerabilities to physical threats (e.g., extreme weather) or cyberattacks. They can also assist in developing response and recovery strategies, complementing planning tools such as the North American Energy Resilience Model. The concept of "GridFMs"—FMs trained on diverse grid data—could significantly advance predictive capabilities, especially for cascading failure scenarios.

Although offering important benefits, FMs also pose risks if misused. The adversarial use of FMs, particularly LLMs, presents significant security risks that can be broadly summarized in two categories: attacks targeting the model itself and attacks leveraging the model as a weapon.

Attacks against the model exploit its vulnerabilities to subvert its intended function or extract sensitive data. This includes prompt injection (or "jailbreaking"), where an attacker crafts input to bypass safety filters and force the model to generate harmful or restricted content. Another major threat is data poisoning, which occurs when malicious data are subtly inserted into the training set, creating hidden backdoors or permanently degrading the model's accuracy. Finally, risks such as model inversion and model stealing compromise confidentiality by allowing adversaries to reconstruct sensitive training data or illegally copy the model's proprietary intelligence.

The second major risk involves using powerful FMs to accelerate and scale traditional cyberattacks. Adversaries leverage these tools to generate highly convincing and personalized phishing e-mails and synthetic media (deepfakes), vastly increasing the success rate of social engineering. FMs also lower the barrier for technical attacks by helping actors write and optimize malicious code or rapidly identify software vulnerabilities, making advanced cyberthreats more common. Furthermore, the complexity of integrating these models into larger systems creates new supply chain risks. For example, a successful prompt injection against an LLM that is integrated with an external tool (i.e., a database) can be used to execute a traditional command injection attack against the connected system, demonstrating that the AI model itself can become a single point of failure and a gateway to broader network compromise.

Users of FMs should invest in AI assurance, red teaming, and development of countermeasures against adversarial applications of FMs, aligning with strategies such as Advance Simulation and Computing's Artificial Intelligence for Nuclear Deterrence program and the Frontiers in Artificial Intelligence for Science, Security and Technology's trustworthy AI pillar.

> **Recommendation 3-3: To address potential security risks arising from the adversarial use of foundation models, the Department of Energy should explore strategies for artificial intelligence assurance, red teaming, and development of countermeasures.**

REFERENCES

Ali, S., P. Arcaini, and A. Arrieta. 2024. "Foundation Models for the Digital Twin Creation of Cyber-Physical Systems." *arXiv*:2407.18779. https://ui.adsabs.harvard.edu/abs/2024arXiv240718779A.

Anysphere. n.d. *Cursor*. https://cursor.com/en, accessed July 31, 2025.

Autodesk. 2025. "Project Bernini: Generative AI 3D Shape Creation." https://www.research.autodesk.com/projects/project-bernini.

Badalassi, V., A. Sircar, J.M. Solberg, J.W. Bae, K. Borowiec, P. Huang, S. Smolentsev, and E. Peterson. 2023. "FERMI: Fusion Energy Reactor Models Integrator." *Fusion Science and Technology* 79(3):345–379.

Benfenati, L., D.J. Pagliari, L. Zanatta, Y.A.B. Velez, A. Acquaviva, M. Poncino, E. MacIi, L. Benini, and A. Burrello. 2025. "Foundation Models for Structural Health Monitoring." *IEEE Transactions on Sustainable Computing*. https://doi.org/10.1109/TSUSC.2025.3592097.

Berger, N., K. Dodge, D.C. Garcia, A. Hughes, L. Kalter, Z. Lu, A. Melfi, et al. 2025. "AI Meets Materials Discovery: The Vision Behind MatterGEN and MatterSIM." *Microsoft Research Blog*. https://www.microsoft.com/en-us/research/story/ai-meets-materials-discovery/?msockid=04adc0effe39670f1736d575ff946616.

Blevins, E.G. 2022. "Global Research and Development Expenditures." Fact Sheet. R44283. https://www.congress.gov/crs-product/R44283#.

Bodnar, C., W.P. Bruinsma, A. Lucic, M. Stanley, A. Allen, J. Brandstetter, P. Garvan, M. Riechert, J.A. Weyn, H. Dong, J.K. Gupta, K. Thambiratnam, A.T. Archibald, C.C. Wu, E. Heider, M. Welling, R.E. Turner, and P. Perdikaris. 2025. "A Foundation Model for the Earth System." *Nature* 641(8065):1180–1187.

Bormon, J.C. 2025. "AI-Assisted Structural Health Monitoring for Foundations and High-Rise Buildings." https://www.preprints.org/manuscript/202509.1196.

Boyce, B.L. 2016. *Born Qualified? The Challenge of Qualifying Additively Manufactured Metals for High-Reliability Applications*. SAND2016-5769PE; 643432. Sandia National Laboratories.

Cao, Z., K. Luo, K. Liu, Y. Cheng, J. Xing, L. Jiang, and J. Fan. 2026. "Physics-Informed Neural Networks for Modeling Turbulent Combustion." *Fuel* 405:136475.

Carey, N., L. Zanisi, S. Pamela, V. Gopakumar, J. Omotani, J. Buchanan, and J. Brandstetter. 2024. "Data Efficiency and Long Term Prediction Capabilities for Neural Operator Surrogate Models of Core and Edge Plasma Codes." *arXiv*:2402.08561. https://ui.adsabs.harvard.edu/abs/2024arXiv240208561C.

Carey, N., L. Zanisi, S. Pamela, V. Gopakumar, J. Omotani, J. Buchanan, J. Brandstetter, F. Paischer, G. Galletti, and P. Setinek. 2025. "Neural Operator Surrogate Models of Plasma Edge Simulations: Feasibility and Data Efficiency." *Nuclear Fusion* 65:106010.

Chan, J., Z. Chen, and E. Pan. 2025. "Foundation Models Knowledge Distillation for Battery Capacity Degradation Forecast." *arXiv*:2505.08151. https://ui.adsabs.harvard.edu/abs/2025arXiv250508151C.

Churchill, R.M. 2024. "AI Foundation Models for Experimental Fusion Tasks." *Frontiers in Physics* 12:531334.

Conti, S. 2024. "Artificial Intelligence for Weather Forecasting." *Nature Reviews Electrical Engineering* 1(1):8.

DOE (Department of Energy). 2024. "Department of Energy Announces $49 Million for Research on Foundational Laboratory Fusion." https://www.energy.gov/science/articles/department-energy-announces-49-million-research-foundational-laboratory-fusion.

DOE. n.d. *Fusion Energy Science*. https://www.energy.gov/science/fes/fusion-energy-sciences, accessed September 29, 2025.

Era, I.Z., I. Ahmed, Z. Liu, and S. Das. 2025. "An Unsupervised Approach Towards Promptable Porosity Segmentation in Laser Powder Bed Fusion by Segment Anything." *npj Advanced Manufacturing* 2(1):10.

Fernandez, J., K.L. Lueth, and P. Wegner. 2025. *Generative AI Market Report 2025–2030*. Market Report. IoT Analytics.

Forien, J.B., G.M. Guss, S.A. Khairallah, W.L. Smith, P.J. DePond, M.J. Matthews, and N.P. Calta. 2023. "Detecting Missing Struts in Metallic Micro-Lattices Using High Speed Melt Pool Thermal Monitoring." *Additive Manufacturing Letters* 4:100112.

Galletti, G., F. Paischer, P. Setinek, W. Hornsby, L. Zanisi, N. Carey, S. Pamela, and J. Brandstetter. 2025. "5D Neural Surrogates for Nonlinear Gyrokinetic Simulations of Plasma Turbulence." *arXiv*:2502.07469. https://ui.adsabs.harvard.edu/abs/2025arXiv250207469G.

Gopakumar, V., S. Pamela, L. Zanisi, Z. Li, A. Anandkumar, and M. Team. 2023. "Fourier Neural Operator for Plasma Modelling." *arXiv*:2302.06542. https://ui.adsabs.harvard.edu/abs/2023arXiv230206542G.

Gottweis, J., and V. Natarajan. 2025. "Accelerating Scientific Breakthroughs with an AI Co-Scientist." *Google Research Blog*. https://research.google/blog/accelerating-scientific-breakthroughs-with-an-ai-co-scientist.

Hassani, S., U. Dackermann, M. Mousavi, and J. Li. 2024. "A Systematic Review of Data Fusion Techniques for Optimized Structural Health Monitoring." *Information Fusion* 103:102136.

Ihme, M., and W.T. Chung. 2024. "Artificial Intelligence as a Catalyst for Combustion Science and Engineering." *Proceedings of the Combustion Institute* 40(1):105730.

Karnakov, P., S. Litvinov, and P. Koumoutsakos. 2024. "Solving Inverse Problems in Physics by Optimizing a Discrete Loss: Fast and Accurate Learning Without Neural Networks." *PNAS Nexus* 3(1):5.

Lam, R., A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, O. Vinyals, J. Stott, A. Pritzel, S. Mohamed, and P. Battaglia. 2023. "Learning Skillful Medium-Range Global Weather Forecasting." *Science* 382(6677):1416–1422.

LLNL (Lawrence Livermore National Laboratory). n.d. "Unique Facilities." https://st.llnl.gov/research/unique-facilities, accessed September 25, 2025.

Martineau, K. 2024. "Meet IBM's New Family of AI Models for Materials Discovery." *IBM Research Blog*. December 20. https://research.ibm.com/blog/foundation-models-for-materials.

McCabe, M., P. Harrington, S. Subramanian, and J. Brown. 2023. "Towards Stability of Autoregressive Neural Operators." *arXiv*:2306.10619. https://ui.adsabs.harvard.edu/abs/2023arXiv230610619M.

McGreivy, N., and A. Hakim. 2024. "Weak Baselines and Reporting Biases Lead to Overoptimism in Machine Learning for Fluid-Related Partial Differential Equations." *Nature Machine Intelligence* 6(10):1256–1269.

Morgan, D., and R. Jacobs. 2020. "Opportunities and Challenges for Machine Learning in Materials Science." *Annual Review of Materials Research* 50:71–103.

NCSES (National Center for Science and Engineering Statistics). 2023. "Survey of Federal Funds for Research and Development: FY 2021–2022." https://ncses.nsf.gov/surveys/federal-funds-research-development/2021-2022.

NVIDIA. n.d. "NVIDIA Omniverse." https://www.nvidia.com/en-us/omniverse, accessed July 31, 2025.

PNNL (Pacific Northwest National Laboratory). n.d. “Fusion Energy Science.” https://www.pnnl.gov/fusion-energy-science, accessed September 29, 2025.

Poels, Y., G. Derks, E. Westerhof, K. Minartz, S. Wiesen, and V. Menkovski. 2023. “Fast Dynamic 1D Simulation of Divertor Plasmas with Neural PDE Surrogates. *Nuclear Fusion* 63:126012.

Poore, C. 2023. “Leveraging Language Models for Fusion Energy Research.” *Princeton Engineering News*. https://engineering.princeton.edu/news/2023/12/20/leveraging-language-models-fusion-energy-research.

Pyzer-Knapp, E.O., M. Manica, P. Staar, L. Morin, P. Ruch, T. Laino, J.R. Smith, and A. Curioni. 2025. “Foundation Models for Materials Discovery—Current State and Future Directions.” *npj Computational Materials* 11(1):61.

Schmude, J., S. Roy, W. Trojak, J. Jakubik, D. Salles Civitarese, S. Singh, J. Kuehnert, et al. 2024. “Prithvi WxC: Foundation Model for Weather and Climate.” *arXiv*:2409.13598. https://ui.adsabs.harvard.edu/abs/2024arXiv240913598S.

Sharma, H., M. Patil, and C. Woolsey. 2020. A Review of Structure-Preserving Numerical Methods for Engineering Applications.” *Computer Methods in Applied Mechanics and Engineering* 366:113067.

Sharma, V., and V. Raman. 2024. “A Reliable Knowledge Processing Framework for Combustion Science Using Foundation Models.” *Energy and AI* 16:100365.

Trewartha, A., N. Walker, H. Huo, S. Lee, K. Cruse, J. Dagdelen, A. Dunn, K.A. Persson, G. Ceder, and A. Jain. 2022. “Quantifying the Advantage of Domain-Specific Pre-Training on Named Entity Recognition Tasks in Materials Science.” *Patterns* 3(4):100488.

Wang, X., S. Liu, A. Tsaris, J.-Y. Choi, A. Aji, M. Fan, W. Zhang, J. Yin, M. Ashfaq, D. Lu, and P. Balaprakash. 2024. “ORBIT: Oak Ridge Base Foundation Model for Earth System Predictability.” *arXiv*:2404.14712. https://ui.adsabs.harvard.edu/abs/2024arXiv240414712W.

Xu, G., M. Jiang, J. Li, X. Xuan, J. Li, T. Lu, and L. Pan. 2024. “Machine Learning-Accelerated Discovery and Design of Electrode Materials and Electrolytes for Lithium Ion Batteries.” *Energy Storage Materials* 72:103710.

Zhang, H., S.D. Semujju, Z. Wang, X. Lv, K. Xu, L. Wu, Y. Jia, et al. 2025. “Large Scale Foundation Models for Intelligent Manufacturing Applications: A Survey.” *Journal of Intelligent Manufacturing*. https://doi.org/10.1007/s10845-024-02536-7.

4

Strategic Considerations and Directions for Department of Energy Foundation Models

Many Department of Energy (DOE) missions demand rapid analysis and decision making under urgent national security or economic constraints. Geopolitical instability can abruptly disrupt access to critical materials essential for defense systems, requiring the swift identification and qualification of substitutes (Dingreville et al. 2024). Shifts in global manufacturing or adversaries' adoption of advanced technologies often force DOE programs to adapt legacy tools and processes to new material systems where empirical data may be scarce and existing models unreliable. Similarly, analysts must forecast the outcomes of nonproliferation or emergency scenarios constrained by complex physical dynamics, such as weather evolution or blast propagation (EoP 2022). These challenges conflict with the traditional trial-and-error discovery cycle that still dominates materials development and qualification. Recent work highlights how data-driven foundation models, integrated with physics-based simulations, can sharply compress these timelines from years to days by guiding targeted experiments and enabling high-fidelity predictions of novel engineered systems (Frey et al. 2025).

The national laboratories hold deep institutional expertise, embedded in their workforce, legacy data sets, and extensive experimental and modeling infrastructure. Yet the sheer scale of the DOE system, characterized by siloed specialized knowledge and the complexity of coordinating a large, distributed workforce, can be fundamentally misaligned with the speed and flexibility required for rapid decision making. Foundation models pose a unique opportunity to automate the coordination of personnel, user facilities and other experimental infrastructure, and historical data to address this long-standing issue of institutional inertia.

Conclusion 4-1: Many DOE missions demand rapid analysis and decision making under urgent national security or economic constraints. While the national laboratories hold deep institutional expertise—embedded in their workforce, legacy data sets, and extensive experimental and modeling infrastructure—the sheer scale of the DOE system, characterized by siloed specialized knowledge and the complexity of coordinating a large, distributed workforce, can be misaligned with the agility required for decisive action. Development of foundation models for this purpose poses a unique opportunity to address rapid analysis and decision making.

Recommendation 4-1: The Department of Energy should explore the use of foundation models to accelerate situational understanding by unifying dispersed, siloed, and diverse multimodal data sources as input to decision-making frameworks across heterogeneous environments.

MATERIAL INFORMATICS AND NEAR-AUTONOMOUS SCIENTIFIC PLATFORMS

In contrast to industrial AI, DOE invested early in material informatics and high-throughput experimental data curation campaigns to build unique access to data sets, through the Material Genome Initiative and other efforts. By combining advanced AI models, high-performance computing, and curated experimental data, materials informatics can dramatically reduce the search space for viable material substitutes or processes. Recent successes demonstrate this potential: for example, generative machine learning approaches have identified candidate alloy systems that reduce dependence on critical rare Earth elements while preserving key performance properties (Dingreville et al. 2024). In another instance, Microsoft researchers screened over 30 million hypothetical compounds to identify new battery cathode chemistries that could cut lithium demand by as much as 70 percent; a discovery pipeline that traditionally would have required years of sequential lab work (Baker 2024). Given DOE's strong software ecosystem, they are uniquely positioned to combine existing efforts where high-throughput fabrication and characterization can be integrated with simulators and knowledge graphs encoding the literature to rapidly identify candidate alternatives for critical materials, processes for manufacturing novel materials, and tools for predicting new materials in poorly understood regimes.

FEDERATED COMPUTING AND DEPARTMENT OF ENERGY FACILITIES

Among DOE's most unique and critical resources are its large-scale user facilities, specialized manufacturing foundries, high-performance computing centers, and shared experimental platforms. Many of these facilities are already equipped with an astronomical number of sensors, generating enormous amounts of data that could be exploited for scientific discovery and process optimization.

Advanced manufacturing facilities such as the Kansas City Plant and Y-12 National Security Complex offer unique opportunities for tailored process improvements if information can be analyzed in a decentralized manner while maintaining necessary controls on classified or sensitive data. The Office of Science has previously invested in federated learning approaches to develop distributed machine learning policies across fleets of assets, including user facilities and other systems, with theoretical guarantees of differential privacy. Related efforts have explored how advanced manufacturing processes, such as metal additive manufacturing, can be coordinated across identical machines operating at multiple sites where local conditions affect performance.

There is now a significant opportunity to integrate these federated systems with foundation models that can process distributed data streams or coordinate physical processes across heterogeneous environments. Such models could take multiple forms: large language models (LLMs) that augment scientists' ability to manage complex distributed systems; agent-based frameworks that execute control policies or distributed data processing; or real-time physics simulators that interpret and contextualize sensor data at scale.

CURATION AND TRANSLATION OF SPECIALIZED KNOWLEDGE

As DOE's workforce turns over, the challenge of maintaining legacy weapons systems and associated hardware or software tools becomes increasingly burdensome; frequently, a single scientist may hold a disproportionate amount of expertise on a given component or system. As staff transition to retirement or alternative career paths, their hard drives may contain vast swaths of data, simulation configuration files, and source code that would take substantial time and financial investment to reproduce. Simultaneously, as new staff are hired, it is broadly understood that there is a steep learning curve to train on the deeply technical software and modeling frameworks used across the laboratories. Foundation models offer a technique to automate the consolidation of existing knowledge and can be used in a copilot configuration to train new members of the workforce, particularly in legacy programming languages or hardware systems that are rarely taught in contemporary university programs.

MULTIMODAL ARTIFICIAL INTELLIGENCE FOR PHYSICS-BASED PREDICTION

Some of the most promising demonstrations of AI-augmented physics simulation have emerged in short-term weather forecasting, where ubiquitous reanalysis data have enabled models that can deliver real-time predictions on a single graphics processing unit, dramatically reducing the computational cost compared to conventional partial differential equation–based solvers at exascale. This creates strategic opportunities to adopt these tools to enhance data-driven decision making and to integrate them into existing physics-based modeling campaigns.

Unique to DOE's mission is the requirement to fuse weather prediction with additional sensing modalities relevant to national security. For example, nonproliferation and counter-terrorism tasks often rely on combining weather models with satellite imagery and other geospatial data. Early industry examples, such as Microsoft's real-time weather foundation models, demonstrate that these models can serve as effective multitasking platforms that generalize well to satellite data streams and other remote sensing tasks.

Beyond this immediate application, the prevalence of diverse scientific data across DOE highlights an opportunity to advance a distinctive form of multimodal learning, extending beyond the text, audio, and video focus common in commercial AI. For example, in stockpile stewardship, it is often necessary to fuse heterogeneous material characterization data—such as X-ray diffraction, electron microscopy, user facility measurements, and high-fidelity simulations—with knowledge graphs and other structured sources, including classified information. Developing foundation models capable of reasoning across such multimodal scientific data streams could establish a unique capability aligned with DOE's national security and scientific missions.

INTEGRATING THE DEPARTMENT OF ENERGY SCIENTIFIC SOFTWARE STACK

While large industrial AI companies have deep expertise in first-order optimizers, automatic differentiation, and other numerical methods central to machine learning, DOE remains a global leader in advanced scientific computing, including large-scale linear algebra; high-performance numerical solvers; higher-order, structure-preserving, and large-scale constrained optimization libraries; and frameworks for discretizing the partial differential equations that underpin scientific simulation. There is a major opportunity to bridge this substantial investment in foundational scientific software with the next generation of foundation models, whether developed by industry or within DOE itself.

As machine learning was initially applied to scientific problems, there was a reluctance within DOE to compete with TensorFlow or PyTorch. At this point, libraries are relatively mature, and open-source libraries such as Trilinos could serve a valuable role in developing lightweight wrapper libraries to facilitate the interfacing of production codes with LLMs. Several notable DOE codes such as MFEM and Albany have begun exposing automatic differentiation and adjoint calculations in a manner that could be accessed by an LLM (MFEM n.d.; Salinger et al. 2016). DOE has invested in higher-level runtime systems that simplify the programming of distributed-memory environments. Frameworks such as Charm++/AMPI (Kale and Krishnan 1993), Legion (Bauer et al. 2012), UPC++ (Bachan et al. 2019), Global Arrays (Nieplocha et al. 1994), and HPX (Kaiser et al. 2014) provide hardware-agnostic abstractions for communication, load balancing, and task scheduling in parallel computing; similar abstractions that facilitate the scheduling of agentic actions or simulation queries for large-scale MPI-style, either as directed by or to build a foundation model, would have value.

A primary function of foundation models is to compress the large corpus into a latent representation that supports multiple downstream tasks. DOE may play a valuable role developing open-source software tools supporting scientific inference from a pretrained latent space. For example, although machine-learned potentials have been widely successful, their implementation within production molecular dynamics simulators such as LAMMPS is often ad hoc, just-in-time–based, and suboptimal in performance. There is a need for a universal library that can distill these classes of data-driven computational kernels into performant, potentially Kokkos-accelerated modules that can be readily deployed in production codes. This opportunity extends beyond LAMMPS to any simulator that would extract data-driven models from a central, pretrained foundation model.

AGENTIC ARTIFICIAL INTELLIGENCE

In the past year, agentic AI has surged as a means of using LLMs to launch external agents to explore hypotheses or improve/verify responses. DOE maintains a collective $407 million per year in open-source code (Shrivastava and Korkmaz 2024), with the Exascale Computing Project alone representing 70 distinct scientific codebases. There is a unique opportunity for DOE to expose automatic differentiation "hooks" in their open-source libraries to allow LLMs to couple directly to production codes, integrating robust numerical prediction into the training process. This would allow LLMs to both perform simulation and calculate loss functions to support holistic end-to-end training through reliable and mature DOE simulators. Several DOE codes already expose adjoints in this manner (see, e.g., MFEM), and so the initial software infrastructure is already in place. In addition to driving simulators in an agentic manner, there is also an opportunity to drive user facilities or autonomous "self-driving" laboratories that generate and process multimodal data. Although multimodal learning is of massive interest to industry, the breadth of modalities, in simulation (ranging from ab initio density functional theory to exascale Earth system models), in experiments (from tabletop X-ray measurements to massive user facilities), and into text (in the form of technical reports and classified journals) dwarfs the more focused efforts likely to be conducted by industry.

Conclusion 4-2: DOE is uniquely positioned to shape the future of AI-driven science. Material informatics and near-autonomous scientific platforms highlight the power of combining curated experimental data, simulation, and advanced AI to accelerate discovery. Federated computing and facility integration extend this vision by enabling distributed use of DOE's infrastructure.

The curation and integration of specialized knowledge coupled with emerging multimodal and agentic AI approaches underscore the importance of preserving expertise, reasoning across diverse scientific data streams, and directly linking foundation models to DOE's mature simulation ecosystem.

Recommendation 4-2: The Department of Energy should both modernize existing infrastructure and invest in new infrastructure to generate, curate, and facilitate the large data corpus necessary to build a scientific foundation model, including simulations to create data, high-throughput and/or autonomous experimental facilities, and facilities to host data. Additionally, they should create interfaces (e.g., agentic, retrieval-augmented generation tools) through which large foundational models may easily access these sources. A successful strategy will provide holistic access to multimodal or heterogeneous infrastructure across the entire DOE complex, mitigating the "stove-piping" of assets between different laboratories or departments.

TALENT RETENTION

The success of any DOE-wide foundation model initiative depends entirely on attracting and retaining top AI talent. This presents significant challenges, primarily due to intense competition from the private sector. Industry has rapidly accelerated its AI hiring, evidenced by a 21 percent increase in AI-related job postings from 2018 to 2023. Critically, employers are now prioritizing practical skill-based hiring over formal degrees. With AI competencies commanding a 23 percent wage premium—a value surpassing that of degrees up to the doctoral level (Bone et al. 2025)—and industry offering higher compensation and exceptional working conditions, DOE will need to compete for this essential expertise.

An added challenge that DOE faces arises from slow funding cycles that make it difficult to keep up with the pace of innovation in industry. Traditional DOE funding cycles, often spanning multiple years, can impede the rapid development and deployment of AI technologies. In contrast, industry laboratories frequently operate with more agile funding mechanisms, enabling quicker adaptation to emerging AI advancements. Within the National Laboratories, laboratory-directed research and development (LDRD)-based funding leads to a minimum

1-year lag to starting a project, which could be slow to the point of missing a major development completely. Furthermore, industry often has the resources to allow teams to solely focus on a single large-scale project, often for long periods on the order of years. To bridge this gap, DOE could consider implementing more flexible funding models, such as rolling proposals or seed grants, to accelerate AI research and development. DOE's Office of Science maintains a number of large multi-institutional initiatives that may provide a vehicle to adapt more flexibly, for example, a Scientific Discovery Through Advanced Computing center, which has a broad enough scope and a sufficiently long-time horizon to adapt to rapid developments in the field while maintaining accountability to taxpayers.

To foster an environment conducive to AI innovation, DOE needs to cultivate a research culture that emphasizes flexibility and speed. This includes adopting performance metrics that prioritize real-world impacts, such as model robustness and deployment success, over traditional academic outputs such as publications. Encouraging interdisciplinary collaboration and providing recognition for contributions to AI systems and infrastructure can further enhance DOE's competitiveness in the AI research landscape.

Despite challenges, DOE possesses unique strengths that can be leveraged to advance AI research and attract talent. DOE engages in mission-driven research; DOE's focus on societal challenges, such as clean energy and national security, attracts scientists motivated by purpose-driven work. Furthermore, in contrast to industry, long-term career tracks within DOE foster sustained development of complex AI systems integrated with physical models. Finally, collaborations between physicists, chemists, computer scientists, and engineers enable the development of AI models that require domain-aware reasoning.

DOE's infrastructure and expertise provide a solid foundation for AI-driven scientific discovery. Decades of investment in physics-based simulation codes offer valuable assets that AI can learn from or emulate. Robust, scalable software platforms developed by DOE laboratories can power hybrid workflows combining symbolic and neural reasoning. Scientific data sets from large-scale experiments serve as high-value training and validation sources for domain-specific AI. Furthermore, DOE's supercomputers and user facilities provide superior computing capabilities and experimental data for training foundation models and deploying AI-augmented simulations.

A further issue is how DOE can best collaborate with universities. Building a strong academic pipeline is crucial for long-term AI capability in DOE. Some possible avenues for encouraging further collaboration with universities include:

- Embedding graduate students and postdocs in national laboratories with co-mentorship from university faculty and lab researchers can strengthen the AI talent pipeline.

- Establishing joint DOE–university institutes focused on the intersection of AI and specific DOE mission areas can foster collaboration and innovation.
- Supporting joint national laboratory and university centers, with potential industry support, that focus on AI and physical sciences can enhance DOE's research capabilities.
- Facilitating easier transitions for university AI experts collaborating with DOE laboratories can promote knowledge exchange and innovation.

Conclusion 4-3: DOE struggles to compete with the private sector for AI talent due to lower salaries and slow, traditional funding cycles. However, DOE's unique strengths, such as its mission-driven work, long-term career paths, and powerful supercomputing infrastructure, can be leveraged to attract talent. Building a strong academic pipeline through closer collaboration with universities is also essential for its long-term success.

Recommendation 4-3: To maintain a top-tier workforce, the Department of Energy (DOE) should design leadership-scale scientific research programs and provide staff with opportunities to rapidly adapt to a quickly evolving technological landscape. To attract early-career scientists, DOE should be perceived as the best place to become a leader in scientific machine learning; while industry may lead large language model space, the unique access to state-of-the-art science can attract top talent. To be competitive with large-scale development efforts in industry, it is important to avoid fracturing of scientists' time and attention. We recommend that DOE should create mechanisms by which medium through large teams can mount coordinated, focused efforts targeting mission-critical developments in fundamental research into, and applications of, foundation models for science.

UNIFIED DATA REPOSITORY

DOE provides several open-source data repositories that serve the research community. These repositories are organized in a fragmented fashion across DOE subdomains (Table 4-1), each hosting heterogeneous data formats and sizes without a unified access interface. Many smaller data sets—often the output of single-investigator LDRD projects—reside on external curation platforms, further fragmenting access. Automated classifiers must inspect each data set for export-control restrictions, adding another layer of procedural complexity. Collected data sets typically represent final project outputs and omit intermediate simulations, classified results, and the metadata and documentation generated during data production.

TABLE 4-1 Department of Energy (DOE) Open-Source Data Repositories

Name	Description
Open Data Catalog	Machine-readable list of all publicly available data sets maintained by DOE and its program and staff offices (https://www.energy.gov/data/articles/open-data-catalog).
DOE Data Explorer (OSTI/E-Link)	Portal for DOE-funded science and engineering data (https://www.osti.gov/dataexplorer).
Materials Data Facility	Publication and discovery service for materials data (Blaiszik et al. 2016; NETL 2024).
Earth System Grid Federation	Archive of climate model output and observations (Ananthakrishnan et al. 2007).
Joint Genome Institute Data Portal	Genomic and metagenomic data sets for bioenergy research (https://data.jgi.doe.gov).
Open Energy Information (OpenEI)	Wiki and repository of energy, resource, and policy data (https://openei.org/wiki/Main_Page).
Wind Integration National Dataset Toolkit	High-resolution wind power meteorology and output data (Draxl et al. 2015).
NREL Data Catalog	Photovoltaic system performance data (https://openei.org/wiki/PVDAQ).

NOTE: NETL = National Energy Technology Laboratory; NREL = National Renewable Energy Laboratory.

DOE can address these challenges by establishing a centralized data center on the scale of its flagship supercomputing facilities. Such a center would offer extensive storage infrastructure, dedicated curation staff, and clear governance policies to enforce a consistent application programming interface for data hosting and retrieval for multimodal scientific data sets. A centralized data center could also help create interfaces not only to access data, but also to access potential foundation models. The easy access to the foundation models could be crucial for the scientific discovery cycle. It would also support research into best practices for data curation and the development of software tools tailored to ingesting large data sets into foundation models.

> *Conclusion 4-4: Although DOE curates many high-value data sets of value for construction of foundation models, they are typically developed in an ad hoc manner with heterogeneous file formats and data curation strategies that currently pose a barrier to high-throughput processing of data. Foundation models present a unique opportunity to address this issue.*

Recommendation 4-4: To increase the success of future foundation models for science, the Department of Energy should invest in large-scale data user facilities (classified and unclassified), leveraged by artificial intelligence's growing capability to interpret heterogeneous scientific data, similar to the successes experienced with previous investments in supercomputers, and open-source scientific computing libraries.

REFERENCES

Ananthakrishnan, R., D.E. Bernholdt, S. Bharathi, D. Brown, M. Chen, A.L. Chervenak, L. Cinquini, et al. 2007. "Building a Global Federation System for Climate Change Research: The Earth System Grid Center for Enabling Technologies (ESG-CET)." *Journal of Physics: Conference Series* 78(1). https://doi.org/10.1088/1742-6596/78/1/012050.

Bachan, J., S. Baden, D. Bonachea, P. Hargrove, S. Hofmeyr, M. Jacquelin, A. Kamil, and B.v. Straalen. 2019. *UPC++ Programmer's Guide, v1.0-2019.3.0*. Tech Report LBNL 2001191. Lawrence Berkeley National Laboratory.

Baker, N. 2024. "Unlocking a New Era for Scientific Discovery with AI: How Microsoft's AI Screened Over 32 Million Candidates to Find a Better Battery." *Microsoft Azure Blog*. https://azure.microsoft.com/en-us/blog/quantum/2024/01/09/unlocking-a-new-era-for-scientific-discovery-with-ai-how-microsofts-ai-screened-over-32-million-candidates-to-find-a-better-battery/?msockid=04adc0effe39670f1736d575ff946616.

Bauer, M., S. Treichler, E. Slaughter, and A. Aiken. 2012. "Legion: Expressing Locality and Independence with Logical Regions." In *SC '12: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*. IEEE. https://doi.org/10.1109/SC.2012.71.

Blaiszik, B., K. Chard, J. Pruyne, R. Ananthakrishnan, S. Tuecke, and I. Foster. 2016. "The Materials Data Facility: Data Services to Advance Materials Science Research." *JOM* 68(8):2045–2052.

Bone, M., E. González Ehlinger, and F. Stephany. 2025. "Skills or Degree?" The Rise of Skill-Based Hiring for AI and Green Jobs." *Technological Forecasting and Social Change* 214:124042.

Dingreville, R., N. Trask, B.L. Boyce, and G.E. Karniadakis. 2024. "Unlocking Alternative Solutions for Critical Materials via Materials Informatics." *The Bridge* 55(2):46–54.

Draxl, C., B.M. Hodge, A. Clifton, and J. McCaa. 2015. *Overview and Meteorological Validation of the Wind Integration National Dataset Toolkit*. Technical Report NREL/TP-5000-61740. National Renewable Energy Laboratory.

EoP (Executive Office of the President of the United States). 2022. "National Strategy for Advanced Manufacturing." https://www.energy.gov/sites/default/files/2024-03/National-Strategy-for-Advanced-Manufacturing-10072022.pdf.

Frey, N.C., I. Hötzel, S.D. Stanton, R. Kelly, R.G. Alberstein, E.Makowski, K. Martinkus, et al. 2025. "Lab-in-the-Loop Therapeutic Antibody Design with Deep Learning." *bioRxiv*: 2025.2002.2019.639050.

Kaiser, H., A. Serio, T. Heller, D. Fey, and B. Adelstein-Lelbach. 2014. "HPX—A Task Based Programming Model in a Global Address Space." In *PGAS '14: Proceedings of the 8th International Conference on Partitioned Global Address Space Programming Models*. https://doi.org/10.1145/2676870.2676883.

Kale, L.V., and S. Krishnan. 1993. "CHARM++: A Portable Concurrent Object Oriented System Based on C++." ACM *SIGPLAN Notices* 28(10):91–108.

MFEM. n.d. "Automatic Differentiation Mini Applications." https://mfem.org/autodiff, accessed September 29, 2025.

NETL (National Energy Technology Laboratory). 2024. “Critical Minerals and Materials Program.” *Program* 141. https://netl.doe.gov/sites/default/files/2024-10/Program-141.pdf.

Nieplocha, J., R.J. Harrison, and R.J. Littlefield. 1994. “Global Arrays: A Portable ‘Shared-Memory’ Programming Model for Distributed Memory Computers.” In *Supercomputing ‘94: Proceedings of the 1994 ACM/IEEE Conference on Supercomputing*, pp. 340–349. IEEE.

Salinger, A.G., R.A. Bartlett, A.M. Bradley, Q. Chen, I.P. Demeshko, X. Gao, G.A. Hansen, et al. 2016. “Albany: Using Component-Based Design to Develop a Flexible, Generic Multiphysics Analysis Code.” *International Journal for Multiscale Computational Engineering* 14(4):415–438.

Shrivastava, R., and G. Korkmaz. 2024. “Measuring Public Open-Source Software in the Federal Government: An Analysis of Code.” *Journal of Data Science* 22(3):356–375.

5

Foundation Model Challenges

SOURCES OF CHALLENGES

Applying foundation models within the Department of Energy's (DOE's) missions presents a multilayered set of technical and operational challenges. These models, which emerged from success in domains such as natural language processing and vision, struggle to transfer directly into DOE's computational science workflows that require physical consistency, mesh- or geometry-aware representations, and scalable inference across high-dimensional, multiscale partial differential equation systems (Pyzer-Knapp et al. 2025). DOE applications such as reactor modeling, Earth systems prediction, and fusion simulation involve high-dimensional, spatiotemporal fields with millions to trillions of values per instance, placing extreme demands on memory, computational throughput, and architectural efficiency. The absence of embedded physical constraints in standard foundation model architectures, combined with stochastic training dynamics, emergent capabilities, and nondeterministic behaviors, hinders scientific reliability, complicates verification, and reduces confidence in high-stakes scenarios (Babuska and Oden 2004). The core promise of foundation models, pretraining across diverse tasks and modalities to enable broad generalization, is precisely what introduces new risks in scientific domains where accuracy, stability, and reproducibility are paramount (Palmer and Stevens 2019). Scientific foundation models are expected to extrapolate across physical regimes, boundary conditions, and domain geometries with minimal adaptation, yet this capability remains largely aspirational in practice. Fine-tuning on downstream scientific problems often proves computationally expensive, brittle, and sensitive to discretization artifacts, with performance degrading when faced with domain shifts or mesh changes (Radova et al. 2025). The lack of standardized data sets for DOE-relevant systems further hampers reproducibility, robust benchmarking, and model transferability across simulation codes or physical domains. DOE use cases also demand task interactivity and feedback integration, such as real-time control of plasma confinement or anomaly detection in sensor networks. These agentic and dynamic workflows are not typically reflected in the static pretraining distributions used to develop generalist foundation models, which are often drawn from web-scale or simulation-agnostic corpora. Consequently, adapting pretrained foundation models for DOE environments requires techniques such as domain-specific simulation environments, reward-informed data-relabeling pipelines, digital twin infrastructure, and architectural modifications that encode physical priors or conservation laws (Yuan et al. 2025). Even in data-rich domains, the absence of reward structures, labeled physics, or causal annotations limits the ability to drive meaningful adaptation. In addition, the need to accommodate heterogeneous data types, such as text, sensor streams, video, and mesh-based simulations, introduces architectural challenges in designing foundation models that can jointly align, fuse, and validate across disparate modalities while preserving spatiotemporal and physical coherence (Mukherjee et al. 2025).

Collaboration with industry introduces additional constraints. Proprietary model weights, restricted data access, and closed-source infrastructure often prevent rigorous verification, validation, and uncertainty quantification (VVUQ) and reproducibility practices, especially when security, transparency, or auditability are required.

Finally, the energy and computational costs of training and adapting large foundation models, particularly across diverse scientific regimes, impose significant burdens on DOE facilities (Koch et al. 2025). Addressing these challenges will require coordinated investments in energy-efficient and sustainable foundation model development, physically informed architectures, domain-specific VVUQ methodologies, and infrastructure for transparent, traceable, and reproducible deployment across DOE's science and national security missions (Teranishi et al. 2025).

Artificial Intelligence Assurance, Test, and Evaluation

AI assurance for foundation models refers to the evidence-based process of demonstrating that a system is reproducible, auditable, and fit for purpose in DOE mission settings. Assurance is tied to acceptance criteria declared in advance for a specific task and operating regime, and results must be repeatable across software environments and hardware platforms. It is not a single evaluation step but a continuous life-cycle discipline spanning model conception, training, deployment, and requalification. This framing echoes emerging life-cycle models for trustworthy AI (Afroogh et al. 2024) and conceptual roadmaps that advocate a "never trust, always verify" paradigm for AI systems (Tidjon and Khomh 2022).

At the requirements stage, DOE programs should specify quantitative criteria for accuracy, stability, and latency. Verification must enforce conformance with physical conservation, invariants, boundary, and unit consistency, and portability across meshes and geometries using both synthetic and experimental data. Validation extends beyond simulation comparison to include closed-loop testing with controllers or optimizers in the loop, where stability margins, constraint-violation rates, and worst-case performance are directly measured. Recent assurance frameworks emphasize that validation should be tied to empirical conditions of use, not only static benchmarks (Bloomfield and Rushby 2024).

Uncertainty quantification is decision linked: predictive coverage should be calibrated against DOE-relevant distributions, and provenance must trace uncertainty sources to modalities, training stages, and preprocessing steps. To support this, foundation models must carry reproducibility dossiers documenting data set lineage, hash-verified snapshots, training seeds, hardware and software stacks, and code commits. Determinism budgets should quantify acceptable drift across multinode and mixed-precision runs. This aligns with recent calls for comprehensive trustworthiness assessment across robustness, transparency, and accountability dimensions (Kowald et al. 2024).

Deployment in high-consequence settings such as fusion control or grid operations requires staged test-beds. Models first undergo software-in-the-loop trials with high-fidelity simulators, advance to hardware-in-the-loop testing on the target control stack, and finally, operate in shadow mode with full telemetry in the live environment. Full deployment proceeds only if predeclared acceptance criteria are satisfied in the simulator and hardware-in-the-loop stages; any modification to data, model, controller, or operating envelope triggers mandatory requalification. This staged life cycle reflects broader proposals for trustworthy and safe AI architecture (Chen et al. 2024) and ensures that DOE's mission applications meet safety and reliability requirements before operational use.

Operational safeguards must be integral to the assurance framework. These include watchdogs and admission control for computing resources, fixed profile execution bounds, and certified fallback controllers. Out-of-distribution detection should be paired with safe degradation policies such as hold-last-good. Where counterfactual reasoning is central, training should be coupled with interventional simulators, and validation should include intervention suites and replay of historical logs. The importance of embedding such protections has been emphasized in the broader AI governance literature (Blau et al. 2024) and in proposals for architectural frameworks for AI safety (Chen et al. 2024).

By consolidating VVUQ, reproducibility, robustness testing, and staged deployment into a unified life cycle, DOE can ensure that foundation models are evaluated with the same rigor long applied to scientific codes.

Verification

Verification ensures that foundation models are implemented correctly and yield outputs consistent with physical principles (Gurieva et al. 2022). For scientific applications at DOE scale, this requires more than standard software testing. The size and complexity of modern foundation models, which often contain billions of parameters and are applied to high-dimensional spatiotemporal fields, demand modular verification strategies that address emergent behaviors, stochastic dynamics, and numerical stability. This is especially critical for systems where any violation of conservation laws or symmetry principles may have safety or operational consequences.

Some DOE applications require inference that is not only accurate but also predictable in timing and auditable in operation so that control and protection functions consistently meet strict deadlines. For these settings, foundation model pipelines must be engineered to satisfy fixed execution budgets, deliver deterministic behavior under load, and fail safely when assumptions are violated. A practical assurance profile includes predictable worst-case execution time established through fixed-profile scheduling on the target platform, hardware-in-the-loop staging before any field activation, and phased deployment that begins in shadow mode with full telemetry before actuation authority is granted. Safety must be ensured through conservative fallback controllers when timing bounds or input validity checks are not met. Continuous audit trails should capture timing, inputs, intermediate states, and actions to ensure full traceability. Additional safeguards include admission control for computing resources, watchdogs, and out-of-distribution input tests that automatically trigger safe states. Acceptance criteria must demonstrate that closed-loop stability and protection margins are preserved across the specified disturbance set and operating envelope. Scientific data sets further complicate the task. Inputs such as three-dimensional mesh-based simulation fields often contain trillions of values, overwhelming conventional memory and computing pipelines. Differences introduced by stochastic initialization, hardware platforms, or software libraries can lead to inconsistent model outputs, undermining reproducibility and making fault tracing difficult (Barton et al. 2022). Most foundation model architectures, especially transformer-based models, are trained on data sets with limited fidelity to physical systems, simulation structure, or simulation-specific structure. As a result, it is difficult to determine whether their predictions honor physical realism, particularly in applications such as turbulent flow or magnetohydrodynamics. These issues are compounded when industry partnerships restrict access to pretrained weights or codebases, limiting transparency and reproducibility (Yang et al. 2020).

Sustainability is another key concern. Verification of large foundation models across multiple scientific domains often involves retraining or revalidation, which incur high energy and computational costs (Han et al. 2023). As model sizes continue to grow, DOE evaluates energy-efficient alternatives and sustainability metrics to ensure that foundation model verification remains viable at scale.

Addressing these challenges requires adopting modular model designs that support isolated testing and interpretation of internal components. This approach is already used in several scientific and engineering pipelines. In operator-learning architectures (Hossain et al. 2025; Kobayashi and Alam 2024; Kobayashi et al. 2025; Lu et al. 2021), the branch-trunk decomposition (e.g., multiple-input operator nets) cleanly separates an encoder branch from a trunk coordinate network, allowing the encoder to be frozen while the trunk is unit-tested on synthetic or gold-standard fields. Neural operator methods with adapter layers create plug-in modules that can be swapped or ablated while holding the core operator architecture fixed (e.g., modular operator learning approaches such as multioperator architecture; Zhang 2024). In hybrid modeling, learned subgrid closures or surrogate modules are routinely inserted into traditional solvers (e.g., turbulence closures in fluid or atmospheric codes) so that the learned module can be validated separately under canonical flow conditions before being integrated into the full solver. (See survey of machine learning closure modeling in turbulence; Beck and Kurz 2021.)

Retrieval-augmented pipelines also already evaluate retriever and predictor modules separately, enabling stress tests of the knowledge interface. Mixture-of-experts (MoE) and routing architectures expose per-expert behavior that can be profiled with targeted inputs and compared against reference cases (e.g., recent MoE gating models showing analyzable expert routing; Nabian and Choudhry 2025).

In practice, isolation is enforced using stable interfaces and test harnesses: strict component contracts for inputs and outputs, synthetic data generators to probe edge-case behavior, golden tests on curated benchmarks, and swap-in or swap-out experiments that leave the surrounding system unchanged except for the module under test. These patterns demonstrate that isolated testing and interpretation are not only possible but already in use in modern scientific and machine learning systems, and they can be extended to foundation models intended for DOE mission-critical deployment.

Benchmarking across DOE high-performance computing (HPC) environments can reduce platform-induced variability, while federated test-beds enable collaboration with industry partners without compromising sensitive intellectual property. Verification efforts should be tightly integrated with comprehensive uncertainty documentation, capturing both aleatory and epistemic components to support robust deployment decisions. To ensure that foundation models are viable for science and engineering, users must treat verification as a foundational component of trust, aligned with sustainability and reproducibility objectives (Mahmood et al. 2024).

Validation

Validation assesses whether foundation model outputs faithfully reflect real-world behavior, particularly in mission-critical DOE applications such as reactor dynamics, grid stability, and materials performance (Wong et al. 2023). This requires systematic comparison of foundation model predictions against experimental observations and high-fidelity simulations (Hsieh et al. 2021), ensuring alignment with physical laws and constraints, such as energy conservation, continuity, and thermodynamic consistency. For complex systems such as microreactors, where safety margins are narrow and data availability is limited, validation must account for data quality, physical plausibility, and generalizability.

High-quality, representative data sets are foundational to foundation model validation. Yet DOE domains often contend with sparse, noisy, or biased data, especially from heterogeneous physical systems such as renewable energy grids or coupled fluid–structure systems. These challenges are compounded by a lack of standardized benchmarks and by the diverse modalities and formats typical of scientific simulations, including scalar, vector, and tensor fields. Furthermore, the geometric dependence of DOE simulations introduces portability concerns, as foundation models trained on one discretization may fail when applied to different meshes or boundary conditions (Brunton et al. 2016; Moscoso et al. 2020).

Validating large-scale, pretrained, multimodal foundation models also entails a significant computational burden. Scientific problems in Earth systems, fusion, or subsurface modeling require validation across spatiotemporal domains and governing equations, often with high-dimensional input–output mappings. Although foundation models are designed to generalize across tasks and scale with data volume, verifying their consistency across multiple physical regimes remains a formidable task (Selin et al. 2024).

To address these issues, DOE can leverage a layered validation strategy. First, experimental cross-validation using real-world data from national user facilities, such as the Advanced Test Reactor, the Advanced Photon Source, or the National Renewable Energy Laboratory, anchors foundation model outputs to physical reality. Second, physics-based benchmarks, such as Monte Carlo neutron transport codes in ExaSMR or SCALE, serve as reference standards for evaluating foundation model fidelity. Where empirical data are sparse, synthetic data sets from validated simulators can support surrogate validation, provided they are curated with traceable metadata and grounded in domain-specific governing equations. For time-critical systems such as fusion control or grid stabilization, validation must also extend to closed-loop behaviors, ensuring stability and performance under uncertainty (Prinn 2013). In turbulence and Earth systems modeling, for example, learned subgrid closures have been validated first on canonical benchmark flows before being integrated into general circulation models, demonstrating that modular surrogate validation is feasible in practice (Beck and Kurz 2021; Hassanian et al. 2025). In nuclear engineering, Monte Carlo neutron transport has long served as a reference standard against which lower-fidelity or surrogate models are calibrated and tested (Leppänen et al. 2015). Similarly, synthetic data from validated simulators are already widely used in fusion and materials science to supplement scarce experimental observations, provided that the synthetic sets carry documented provenance and are grounded in governing equations (Kobayashi et al. 2025). Recent surrogate modeling studies further reinforce this layered approach, including climate emulation with graph neural networks (Potter et al. 2024), coastal ocean circulation surrogates with physics-based constraints (Xu et al. 2024), adaptive implicit neural representations for high-fidelity scientific simulations (Li et al. 2025), and surrogate-based Bayesian calibration frameworks for climate models (Holthuijzen et al. 2025). Mesh portability challenges have been addressed using graph neural network surrogates on unstructured grids (Shi et al. 2022), and DOE's Oak Ridge National Laboratory has employed surrogate-based calibration of the E3SM atmosphere model (Yarger et al. 2024). Physics-informed surrogate models have also been demonstrated for groundwater transport forecasting (Meray et al. 2024), while diffusion-based surrogates are emerging for regional climate and sea-ice simulations (Finn et al. 2024). These precedents indicate that the layered validation strategy is not speculative but reflects a growing body of practice across multiple domains.

Importantly, validation is not a binary pass/fail exercise. If a foundation model is shown to be invalid for a given regime, it is not discarded wholesale; instead, its use is confined to conditions where validation evidence is sufficient. In DOE mission settings, this means restricting the model to advisory or exploratory roles until retraining, fine-tuning, or hybridization with physics solvers restores fidelity. Models may also be demoted to shadow-mode operation, where outputs are logged but not acted upon until requalification criteria are met. This mirrors the way traditional simulation codes undergo continuous VVUQ cycles rather than one-time certification. Thus, the layered validation framework both builds on prior evidence and provides structured pathways for handling failure, ensuring that only models with verified domain fidelity are elevated to operational use.

Uncertainty Quantification

Uncertainty Quantification (UQ) is indispensable for the trustworthy use of foundation models in DOE applications (Bilbrey et al. 2025). Unlike traditional simulators with interpretable inputs and outputs, foundation models pretrained on diverse tasks and modalities behave as black box approximators whose outputs are not explicitly governed by physical laws. This creates deep challenges for UQ, as error sources can propagate across input types, scientific contexts, or temporal regimes without clear attribution or traceability (Wang et al. 2023). Validation cannot rely on predictive fit alone when DOE decisions depend on counterfactuals and operator interventions. Foundation models must preserve causal structure under changes in operating point, control actions, and boundary conditions. Meeting this challenge requires integrating causal formalisms and intervention-based testing into both training and validation. Practical approaches include incorporating physics-based causal graphs or invariance penalties during training, pairing learning with interventional simulators that generate policy-relevant counterfactuals, and extending validation to intervention suites derived from simulation campaigns and historical logs. Evidence of robustness should include not just predictive accuracy but counterfactual fidelity, invariance under admissible interventions, and stability when the model is exercised in closed-loop control settings. Recognizing causal and interventional robustness as a distinct challenge ensures that DOE foundation models are capable of supporting decision making in safety-critical and policy-relevant environments.

Pretrained foundation models used in DOE settings must often integrate sparse, noisy, or out-of-distribution data to support scientific inference (Moro et al. 2025). This introduces layered uncertainties: aleatory uncertainty from inherent randomness, epistemic uncertainty from incomplete knowledge, and structural uncertainty arising due to domain shift between pretraining and deployment (Moscoso et al. 2020). For example, a foundation model trained on geophysical sensor networks may fail to generalize to grid control scenarios if rare but critical events are underrepresented. Without explicit UQ, narrow predictive intervals may mask failure risks that compromise safety and mission assurance. Multimodal foundation models compound this complexity. Architectures that integrate text, telemetry, simulation output, and high-resolution spatiotemporal fields confront alignment and calibration issues unique to each data type. Classical UQ techniques, which assume homogeneity of inputs and well-defined likelihoods, are poorly suited to these heterogeneous scientific settings. Pretraining on unlabeled corpora also introduces ambiguity about data provenance, fidelity, and representativeness, weakening the basis for uncertainty estimation in downstream DOE applications.

DOE applications demand not only accurate predictions but transparent characterization of uncertainty across heterogeneous data sources. Foundation models must therefore estimate and report uncertainty per modality before composing it at the task level. Each input class, whether text, point sensors, images, or simulation fields, requires its own calibrated noise model and uncertainty head, with ensembles or Bayesian layers providing epistemic estimates of model uncertainty. Out-of-distribution detection should operate at both the single-modality and joint levels to flag inputs outside training distributions. Coverage guarantees must be calibrated with conformal or likelihood-free methods on DOE-relevant distributions to ensure reliability. Every prediction should carry a structured uncertainty record that attributes contributions to specific modalities, training stages, and preprocessing steps. Such provenance enables users, operators, and regulators to understand not only the magnitude of uncertainty but its origin, providing the transparency required for deployment in high-consequence DOE missions. For DOE's mission-critical use, uncertainty must not only be quantified but also interpretable to domain experts and regulators (NEA 2016). While ensemble methods and Bayesian deep learning offer statistical tools, they do not fully meet DOE's high-dimensional and context-sensitive requirements (Fort et al. 2020). In domains such as fusion energy or nuclear thermal hydraulics, UQ must resolve sensitivity to mesh discretization, boundary geometry, and initial condition variability (Wang et al. 2022). UQ must be integrated into foundation model pipelines from the outset, rather than retrofitted postdeployment. Early inclusion allows recursive calibration, scenario-based testing, and adaptive trust assessment as models are transferred across domains or facility environments.

For DOE's mission-critical settings, predictive fit alone is insufficient. Decision support often requires counterfactual reasoning: how a system responds under interventions such as operator actions, set-point changes, or equipment faults. Foundation models must therefore be validated not just on observed data but on their behavior under interventions and in closed-loop interaction with controllers. Integrating causal robustness into DOE's assurance framework requires physics-informed causal graphs or invariance penalties during training, coupled with interventional simulators that generate policy-relevant counterfactuals. Validation should extend to structured intervention suites built from simulations and historical logs. Pairing uncertainty quantification with causal checks during pretraining and fine-tuning enables early rejection of models that may replicate passively observed data but collapse under perturbation. Evidence of robustness must include counterfactual fidelity, invariance under admissible interventions, and stability when embedded in control loops. Recognizing causal and interventional robustness as a distinct challenge ensures that foundation models can support DOE operators and regulators with trustworthy, decision-relevant behavior. This alignment with validation and reproducibility workflows gives DOE decision makers a reliable basis for quantifying and managing uncertainty in operational systems (Rudin 2019), with test-beds such as DOE's Nuclear Energy Advanced Modeling and Simulation program (NEAMS n.d.) and Office of Cybersecurity, Energy Security, and Emergency Response (DOE n.d.) offering structured platforms for future foundation model–UQ integration).

Ultimately, general-purpose foundation models are not viable for deployment in DOE's regulatory and high-risk environments without multimodal, physics-aware, and domain-transferable UQ mechanisms that match the complexity and societal stakes of DOE science. Although foundation models offer compelling new capabilities, DOE cannot assume that existing VVUQ practices for traditional simulation codes apply directly. At present, foundation models should be pursued as research assets whose deployment in high-consequence settings depends on the creation of assurance frameworks. This means that near-term use is appropriate for exploratory science, surrogate modeling, and advisory applications, but operational roles in control, protection, or licensing should await the development of DOE-specific VVUQ, reproducibility, and assurance standards. Thus, the immediate recommendation is not to prohibit use but to invest in dedicated research that adapts and extends VVUQ methods to the foundation model context, establishing the evidence base required for safe and certifiable deployment.

> *Conclusion 5-1: VVUQ methods analogous to those for traditional computational modeling do not exist for, or map directly onto, foundation models.*

REPRODUCIBILITY

Reproducibility is the ability to replicate results under consistent conditions, a foundational requirement for scientific integrity and model trustworthiness. In the context of foundation models, especially those pretrained across heterogeneous data modalities and designed for cross-task generalization, reproducibility becomes significantly more complex. These models are often trained on massive, uncurated data sets, under evolving software environments and stochastic training routines (Laine et al. 2021). Such variability introduces silent failure modes that can undermine reliability in DOE's high-stakes domains (Tian et al. 2018), where model outputs may influence nuclear safety evaluations, advanced material qualification, or infrastructure resilience planning (Wang et al. 2025). Unlike narrowly scoped machine learning models, foundation models function as multipurpose, continuously evolving systems. Their ability to generalize across modalities (e.g., text, simulation data, and sensor fields) and across tasks introduces deeper reproducibility risks. The same model may be applied to subchannel thermal-hydraulics in one instance and to geospatial risk mapping in another, with minimal retraining. Without rigorous documentation of pretraining data sets, transfer learning decisions, and model evolution, the provenance of any single prediction becomes difficult to verify or audit. Moreover, generalist models often operate with latent knowledge acquired during pretraining stages that are difficult to retrace or validate (Pyzer-Knapp et al. 2025).

In DOE contexts, these concerns are not academic. Reproducibility is a precondition for regulatory acceptance, operational deployment, and scientific validation (Allison et al. 2018). Yet, three critical barriers persist. First, nondeterminism due to random weight initialization, floating-point discrepancies, and hardware variability can yield different outputs for the same inputs, especially when dealing with distributed training across heterogeneous platforms (Allison et al. 2018). Second, data and code access are often restricted in national security or proprietary collaborations, making external replication difficult. Third, inconsistent training practices (e.g., undocumented hyperparameters, varying data preprocessing pipelines, or ad hoc fine-tuning) introduce methodological drift across teams and institutions (Nichols et al. 2021).

Addressing these challenges requires intentional infrastructure and cultural shifts. Standardized computing environments, reproducible pipelines using fixed seeds and version-controlled dependencies, and MLOps tooling for experiment lineage must become baseline practices (Nature.com 2021). DOE is uniquely positioned to lead here, leveraging its HPC systems and scientific workflow infrastructure to enforce deterministic model training and versioned data sets. Open science policies, where feasible, should promote model card documentation, training log archival, and reproducibility benchmarks. In secure settings, controlled-access reproducibility testbeds can support internal verification without exposing sensitive materials. Ultimately, the reproducibility of foundation models in science depends on shared codebases, fixed sources of randomness, and acknowledging that foundation models are not static endpoints but evolving, reusable artifacts (Nichols et al. 2021). Reproducibility must account for how a model was trained on what data, for which task, and under which assumptions, while enabling traceable, auditable reuse across new applications. This becomes especially vital as DOE seeks to deploy general-purpose models across institutions and missions, where latent variability may propagate unnoticed and compromise reliability at scale.

In DOE contexts, *fit-for-purpose* means that a foundation model can be demonstrated to satisfy acceptance criteria that are explicitly matched to the safety, security, and reliability demands of its intended use. For exploratory science and low-risk applications, this may require only statistical fidelity, convergence under refinement, and reproducibility of results across runs and platforms. For regulatory or mission-relevant applications, fit-for-purpose raises the bar: models must provide deterministic behavior within specified tolerances, complete provenance of data and training decisions, and calibrated uncertainty estimates with coverage guarantees tied to DOE-relevant distributions. For real-time control or protection functions, fit for purpose requires safety certification: predictable execution under bounded latency and jitter, validated closed-loop stability margins, and robust fallback or fail-safe behavior under disturbance.

Mapping VVUQ to these tiers ensures that DOE foundation models are not treated as "one size fits all," but are qualified according to the risks they manage. Tiered acceptance criteria might include (1) reproducibility benchmarks and physics-based consistency checks for discovery science; (2) reproducibility dossiers, provenance logging, and validated uncertainty quantification for regulatory use; and (3) hardware-in-the-loop timing guarantees, interventional validation suites, and documented fail-safe policies for mission-critical control. By embedding these criteria, fit for purpose becomes an operational standard rather than a rhetorical goal, aligning model trustworthiness with the concrete safety, security, and reliability needs of DOE missions.

Conclusion 5-2: VVUQ, interpretability, and reproducibility are critical for establishing and maintaining trust in systems that are inherently complex, opaque, and increasingly deployed in high-stakes situations. Integration of VVUQ into foundation models would lead to increasing their trustworthiness, reliability, and fit for purpose, which is essential for future scientific discovery and innovation.

Recommendation 5-1: The Department of Energy (DOE) should lead the development of verification, validation, and uncertainty quantification frameworks tailored to foundation models, with built-in support for physical consistency, structured uncertainty quantification, and reproducible benchmarking in DOE-relevant settings.

Conclusion 5-3: AI for science will demand more and different physical experiments to validate the veracity of the AI predictions. Empirical grounding ensures that foundation model outputs reflect physical laws and domain-specific behavior. This is especially critical in high-stake DOE applications, where simulations alone cannot guarantee correctness, and where physical experiments provide the only definitive test of predictive validity.

Recommendation 5-2: In line with Recommendation 4-2, the Department of Energy should place high priority on data collection efforts to support reproducible foundation model training and validation, analogous to traditional efforts in verification, validation, and uncertainty quantification.

Recommendation 5-3: The Department of Energy should establish and enforce standardized protocols and develop benchmarks for training, documenting, and reproducing foundation models for science and should participate in defining software standards, addressing randomness, hardware variability, and data access across its laboratories and high-performance computing infrastructure.

CHALLENGES OF INDUSTRIAL COLLABORATION

There are both benefits and risks when collaborating with AI industry leaders. It would benefit DOE to be aware of such benefits and the challenges that collaboration might bring.

- *Benefits and risks:* Industrial partnerships provide DOE with access to advanced computational platforms, specialized foundation model expertise, and scalable software pipelines, accelerating the development and deployment of foundation models. A notable example is the Pacific Northwest National Laboratory (PNNL)–Microsoft collaboration, which leveraged AI and HPC to identify a solid-state electrolyte that reduced lithium usage significantly (*ScienceAdviser* 2024). This collaboration exemplifies the potential of combining domain science with state-of-the-art industrial infrastructure. However, such partnerships introduce risks, including restricted access to training data and model weights, proprietary architectures, and diverging priorities, as industry tends to emphasize market-driven goals. DOE, by contrast, upholds a public science and national security mission.
- *Proprietary technology and data sharing:* Proprietary models and data sets can inhibit VVUQ and reproducibility (Bail 2024). DOE projects involving classified or legacy data face additional barriers in adopting or

modifying commercial foundation models. Licensing terms, intellectual property concerns, and export controls necessitate structured agreements. Techniques such as federated learning or secure APIs can mitigate exposure risks but introduce technical and coordination burdens. The PNNL–Microsoft case illustrates the need for structured interfaces that advance science without compromising data integrity or transparency.

- *Balancing commercial and domain-specific models:* The trade-off between using commercial foundation models (e.g., GPT-5) and developing domain-specific models tailored to DOE needs is not trivial. Commercial models are often multimodal and efficient but may underperform in accuracy-critical settings such as reactor kinetics or plasma control (Sarker 2022). By contrast, domain-specific models align better with physical constraints but require significant DOE investment in data curation, model training, and infrastructure. Hybrid strategies such as fine-tuning open-source backbones, incorporating retrieval-based augmentation, or adopting tiered licensing can help DOE benefit from commercial models while retaining control over mission-sensitive functionality. Recent studies show that commercial foundation models can provide valuable starting points for DOE use when carefully adapted. For example, large language models pretrained on general corpora have been successfully fine-tuned for domain science tasks such as materials property prediction, protein folding, and scientific code generation. In Earth sciences, general vision–language models have been adapted to remote sensing and climate data through retrieval-augmented pipelines, significantly reducing the cost of training from scratch. Hybrid strategies that combine open-weight commercial backbones with DOE-curated data have already demonstrated improved sample efficiency and reduced infrastructure costs compared to fully bespoke models. These precedents indicate that DOE can benefit from commercial models not by adopting them wholesale, but by treating them as adaptable baselines that lower entry costs and accelerate deployment while preserving pathways for domain-specific fine-tuning and assurance.
- *Computational and data infrastructure:* Cloud-based industrial infrastructure enables scalable model training and inference but raises concerns regarding sustained access, reproducibility, and dependence on vendor-controlled platforms (Talirz et al. 2020). DOE workflows often rely on legacy simulation pipelines and experimental tools, raising interoperability challenges when coupled with commercial AI ecosystems. Data curation remains a core barrier, especially for multimodal pipelines combining sensor data, structured simulations, and annotated experimental data sets. The energy intensity of foundation model operations also demands green computing strategies and life cycle–aware efficiency metrics.

- *Ethical considerations:* Partnerships must be structured to uphold ethical integrity. Commercial foundation models may reflect biases from pretraining corpora or behave unpredictably in edge-case scientific scenarios (Blau et al. 2024). For DOE, safeguarding sensitive data, ensuring equitable outputs, and protecting scientific independence are paramount. Governance mechanisms should enforce bias auditing, usage transparency, and responsible development aligned with DOE's public mission.

ETHICS, SAFETY, AND GOVERNANCE

The following measures frame responsible use and deployment alongside validation, verification, and uncertainty quantification.

- *Dual use and misuse:* Foundation models designed for DOE science may be repurposed in unintended ways, including adversarial cyber operations, weaponization of scientific knowledge, or unauthorized manipulation of critical infrastructure. The dual-use dilemma is acute when models trained on sensitive nuclear, grid, or materials data are shared without safeguards. Addressing this challenge requires clear DOE policies on access control, responsible licensing, and the use of model cards that specify intended scope and restrictions. Technical safeguards should include purpose-binding at the workflow or application programming interface level, filters that block disallowed prompts or inference chains, and approval gates for sensitive features. Usage must be logged with auditable traces and rate limits, while misuse red-teaming and rollback procedures are incorporated into routine evaluation cycles.
- *Equity and bias in scientific data:* Training data drawn from scientific facilities, simulations, or environmental sensors may contain geographic, demographic, or institutional biases that propagate into downstream analyses. For instance, models trained primarily on data from well-instrumented regions may underperform in underserved or developing contexts, reinforcing inequities. To mitigate these risks, DOE foundation model pipelines should embed bias-aware curation practices such as stratified sampling, augmentation of underrepresented regimes, and per-modality calibration. Coverage maps can identify blind spots, while model cards disclose data composition, known biases, and intended scope so that downstream users avoid unsupported applications.
- *Safety of autonomous lab actions:* Foundation models integrated into experimental workflows, robotics, or closed-loop laboratories introduce new safety hazards. Mis-specified objectives, misinterpreted sensor inputs, or adversarial perturbations could lead to unsafe behavior in laboratories handling hazardous materials or operating advanced reactors. Assurance mechanisms must include explicit interlocks, real-time human

supervision, and shadow-mode testing before any autonomous authority is granted. Embedding these safeguards ensures that DOE facilities can benefit from automation while avoiding catastrophic failures driven by misaligned model behavior.

- *Provenance and accountability:* Provenance is critical for ensuring that predictions, recommendations, or control actions can be traced back through pretraining data, fine-tuning procedures, and deployment environments. Without auditable lineage, regulators and operators cannot verify whether outputs meet DOE's trust and safety thresholds. Meeting this challenge requires reproducibility dossiers and audit trails that capture versioned data sets, training seeds, software environments, and intervention histories. Hardware and environment profiles should be logged, with signatures or attestations verifying workflow identity. This infrastructure enables reproducibility reviews, external audits, and proper attribution across institutions.
- *Energy and sustainability accounting:* Training and retraining large-scale foundation models consume significant energy, sometimes on the scale of DOE's HPC facilities. Sustainability must therefore become a first-class dimension of assurance. DOE should require reporting of energy per training run and per inference, prioritize compact adapters and retrieval methods over full retraining when possible, and schedule large jobs to align with cleaner energy windows where feasible. Hardware selection should emphasize meeting latency requirements at the lowest practical power cost. By embedding sustainability metrics into VVUQ frameworks, DOE can ensure that AI deployment advances a reliable, affordable, and clean energy future in line with its mission.

Conclusion 5-4: Partnering of DOE laboratories with industry on AI foundation models will require deliberate effort, including flexible contracting mechanisms, clear intellectual property agreements, data-sharing processes, aligning on VVUQ approaches, responsible AI practices, and a shared understanding of respective missions, objectives, and constraints.

Recommendation 5-4: The Department of Energy should deliberately pursue partnerships with industry and academia to address national mission goals, governed by flexible contracts, responsible artificial intelligence standards, and alignment on reproducibility, verification, validation, and uncertainty quantification approaches and data sharing.

REFERENCES

Afroogh, S., A. Akbari, E. Malone, M. Kargar, and H. Alambeigi. 2024. "Trust in AI: Progress, Challenges, and Future Directions." *arXiv*:2403.14680. https://ui.adsabs.harvard.edu/abs/2024arXiv240314680A.

Allison, D.B., R.M. Shiffrin, and V. Stodden. 2018. "Reproducibility of Research: Issues and Proposed Remedies." *Proceedings of the National Academy of Sciences of the United States of America* 115(11):2561–2562.

Babuska, I., and J.T. Oden. 2004. "Verification and Validation in Computational Engineering and Science: Basic Concepts." *Computer Methods in Applied Mechanics and Engineering* 193(36–38): 4057–4066.

Bail, C.A. 2024. "Can Generative AI Improve Social Science?" *Proceedings of the National Academy of Sciences of the United States of America* 121(21):e2314021121.

Barton, C.M., A. Lee, M.A. Janssen, S. van der Leeuw, G.E. Tucker, C. Porter, J. Greenberg, et al. 2022. "How to Make Models More Useful." *Proceedings of the National Academy of Sciences of the United States of America* 119(35):e2202112119.

Beck, A., and M. Kurz. 2021. "A Perspective on Machine Learning Methods in Turbulence Modeling." *GAMM Mitteilungen* 44(1):e202100002.

Bilbrey, J.A., J.S. Firoz, M.S. Lee, and S. Choudhury. 2025. "Uncertainty Quantification for Neural Network Potential Foundation Models." *npj Computational Materials* 11(1):109.

Blau, W., V.G. Cerf, J. Enriquez, J.S. Francisco, U. Gasser, M.L. Gray, M. Greaves, et al. 2024. "Protecting Scientific Integrity in an Age of Generative AI." *Proceedings of the National Academy of Sciences of the United States of America* 121(22):e2407886121.

Bloomfield, R., and J. Rushby. 2024. "Assurance of AI Systems from a Dependability Perspective." *arXiv*:2407.13948. https://ui.adsabs.harvard.edu/abs/2024arXiv240713948B.

Brunton, S.L., J.L. Proctor, and J.N. Kutz. 2016. "Discovering Governing Equations from Data by Sparse Identification of Nonlinear Dynamical Systems." *Proceedings of the National Academy of Sciences of the United States of America* 113(15):3932–3937.

Chen, C., X. Gong, Z. Liu, W. Jiang, S.Q. Goh, and K.-Y. Lam. 2024. "Trustworthy, Responsible, and Safe AI: A Comprehensive Architectural Framework for AI Safety with Challenges and Mitigations." *arXiv*:2408.12935. https://ui.adsabs.harvard.edu/abs/2024arXiv240812935C.

DOE (Department of Energy). n.d. "About the Office of Cybersecurity, Energy Security, and Emergency Response (CESER)." https://www.energy.gov/ceser/about-office-cybersecurity-energy-security-and-emergency-response, accessed July 31, 2025.

Finn, T.S., C. Durand, A. Farchi, M. Bocquet, P. Rampal, and A. Carrassi. 2024. "Generative Diffusion for Regional Surrogate Models from Sea-Ice Simulations." *Journal of Advances in Modeling Earth Systems* 16:e2024MS004395.

Fort, S., H. Hu, and B. Lakshminarayanan. 2020. "Deep Ensembles: A Loss Landscape Perspective." *ArXiv*. https://doi.org/10.48550/arXiv.1912.02757.

Gurieva, J., E. Vasiliev, and L. Smirnov. 2022. "Application of Conservation Laws to the Learning of Physics-Informed Neural Networks." *Procedia Computer Science* 212:464–473.

Han, B.A., K.R. Varshney, S. LaDeau, A. Subramaniam, K.C. Weathers, and J. Zwart. 2023. "A Synergistic Future for AI and Ecology." *Proceedings of the National Academy of Sciences of the United States of America* 120(38):e2220283120.

Hassanian, R., Á. Helgadóttir, F. Gharibi, A. Beck, and M. Riedel. 2025. "Data-Driven Deep Learning Models in Particle-Laden Turbulent Flow." *Physics of Fluids* 37(2):023348.

Holthuijzen, M.F., A. Chakraborty, E. Krath, and T. Catanach. 2025. "Surrogate-Based Bayesian Calibration Methods for Climate Models: A Comparison of Traditional and Non-Traditional Approaches." *arXiv*:2508.13071. https://ui.adsabs.harvard.edu/abs/2025arXiv250813071H.

Hossain, R., F. Ahmed, K. Kobayashi, S. Koric, D. Abueidda, and S.B. Alam. 2025. "Virtual Sensing-Enabled Digital Twin Framework for Real-Time Monitoring of Nuclear Systems Leveraging Deep Neural Operators." *npj Materials Degradation* 9(1):21.

Hsieh, A.S., K.A. Brown, N.B. deVelder, T.G. Herges, R.C. Knaus, P.J. Sakievich, L.C. Cheung, B.C. Houchens, M.L. Blaylock, and D.C. Maniaci. 2021. "High-Fidelity Wind Farm Simulation Methodology with Experimental Validation." *Journal of Wind Engineering and Industrial Aerodynamics* 218:104754.

Kobayashi, K., and S.B. Alam. 2024. "Deep Neural Operator-Driven Real-Time Inference to Enable Digital Twin Solutions for Nuclear Energy Systems." *Scientific Reports* 14(1):2101.

Kobayashi, K., S. Roy, S. Koric, D. Abueidda, and S. Bahauddin Alam. 2025. "From Proxies to Fields: Spatiotemporal Reconstruction of Global Radiation from Sparse Sensor Sequences." *arXiv*:2506.12045. https://ui.adsabs.harvard.edu/abs/2025arXiv250612045K.

Koch, F., A. Djuhera, and A. Binotto. 2025. "Intelligent Orchestration for Inference of Large Foundation Models at the Edge." *arXiv*. https://doi.org/10.48550/arXiv.2504.03668.

Kowald, D., S. Scher, V. Pammer-Schindler, P. Müllner, K. Waxnegger, L. Demelius, A. Fessl, et al. 2024. "Establishing and Evaluating Trustworthy AI: Overview and Research Challenges." *arXiv*:2411.09973. https://ui.adsabs.harvard.edu/abs/2024arXiv241109973K.

Laine, R.F., I. Arganda-Carreras, R. Henriques, and G. Jacquemet. 2021. "Avoiding a Replication Crisis in Deep-Learning-Based Bioimage Analysis." *Nature Methods* 18(10):1136–1144.

Leppänen, J., M. Pusa, T. Viitanen, V. Valtavirta, and T. Kaltiaisenaho. 2015. "The Serpent Monte Carlo Code: Status, Development and Applications in 2013." *Annals of Nuclear Energy* 82: 142–250.

Li, Z., Y. Duan, T. Xiong, Y.-T. Chen, W.-L. Chao, and H.-W. Shen. 2025. "High-Fidelity Scientific Simulation Surrogates via Adaptive Implicit Neural Representations." *arXiv*:2506.06858. https://ui.adsabs.harvard.edu/abs/2025arXiv250606858L.

Lu, L., P. Jin, G. Pang, Z. Zhang, and G.E. Karniadakis. 2021. "Learning Nonlinear Operators via DeepONet Based on the Universal Approximation Theorem of Operators." *Nature Machine Intelligence* 3(3):218–229.

Mahmood, S., H. Sun, A.A. Alhussan, A. Iqbal, and E.-S.M. El-kenawy. 2024. "Active Learning-Based Machine Learning Approach for Enhancing Environmental Sustainability in Green Building Energy Consumption." *Scientific Reports* 14(1):19894.

Meray, A., L. Wang, T. Kurihana, I. Mastilovic, S. Praveen, Z. Xu, M. Memarzadeh, A. Lavin, and H. Wainwright. 2024. "Physics-Informed Surrogate Modeling for Supporting Climate Resilience at Groundwater Contamination Sites." *Computers & Geosciences* 183:105508.

Moro, V., C. Loh, R. Dangovski, A. Ghorashi, A. Ma, Z. Chen, S. Kim, P.Y. Lu, T. Christensen, and M. Soljačić. 2025. "Multimodal Foundation Models for Material Property Prediction and Discovery." *Newton* 1(1):100016.

Moscoso, M., A. Novikov, G. Papanicolaou, and C. Tsogka. 2020. "The Noise Collector for Sparse Recovery in High Dimensions." *Proceedings of the National Academy of Sciences of the United States of America* 117(21):11226–11232.

Mukherjee, S., J. Lang, O. Kwon, I. Zenyuk, V. Brogden, A. Weber, and D. Ushizima. 2025. "Foundation Models for Zero-Shot Segmentation of Scientific Images Without AI-Ready Data." *arXiv Computer Science*. https://doi.org/10.48550/arXiv.2506.24039.

Nabian, M.A., and S. Choudhry. 2025. "A Mixture of Experts Gating Network for Enhanced Surrogate Modeling in External Aerodynamics." *arXiv*:2508.21249. https://ui.adsabs.harvard.edu/abs/2025arXiv250821249N.

Nature.com. 2021. "Moving Towards Reproducible Machine Learning." *Nature Computational Science* 1(10):629–630.

NEA (Nuclear Energy Agency). 2016. *Review of Uncertainty Methods for Computational Fluid Dynamics Application to Nuclear Reactor Thermal Hydraulics*. Organisation for Economic Co-operation and Development.

NEAMS (Nuclear Energy Advanced Modeling and Simulation). n.d. *About*. https://neams.inl.gov/about-us, accessed July 31, 2025.

Nichols, J.D., M.K. Oli, W.L. Kendall, and G.S. Boomer. 2021. "A Better Approach for Dealing with Reproducibility and Replicability in Science." *Proceedings of the National Academy of Sciences of the United States of America* 118(7):e2100769118.

Palmer, T., and B. Stevens. 2019. "The Scientific Challenge of Understanding and Estimating Climate Change." *Proceedings of the National Academy of Sciences of the United States of America* 116(49):34390–34395.

Potter, K., C. Martinez, R. Pradhan, S. Brozak, S. Sleder, and L. Wheeler. 2024. "Graph Convolutional Neural Networks as Surrogate Models for Climate Simulation." *arXiv*:2409.12815. https://ui.adsabs.harvard.edu/abs/2024arXiv240912815P.

Prinn, R.G. 2013. "Development and Application of Earth System Models." *Proceedings of the National Academy of Sciences of the United States of America* 110(Suppl. 1):3673–3680.

Pyzer-Knapp, E.O., M. Manica, P. Staar, L. Morin, P. Ruch, T. Laino, J.R. Smith, and A. Curioni. 2025. "Foundation Models for Materials Discovery—Current State and Future Directions." *npj Computational Materials* 11(1):61.

Radova, M., W.G. Stark, C.S. Allen, R.J. Maurer, and A.P. Bartók. 2025. "Fine-Tuning Foundation Models of Materials Interatomic Potentials with Frozen Transfer Learning." *npj Computational Materials* 11(1):237.

Rudin, C. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1(5):206–215.

Sarker, I.H. 2022. "AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems." *SN Computer Science* 3(2):158.

ScienceAdviser. 2024. "Accelerating the Discovery of Battery Materials with AI." https://www.science.org/do/10.1126/science.zw4uuid/full/_20240216_cpub_microsoft_feature-1714408300127.pdf.

Selin, N.E., A. Giang, and W.C. Clark. 2024. "Showcasing Advances and Building Community in Modeling for Sustainability." *Proceedings of the National Academy of Sciences of the United States of America* 121(29):e2215689121.

Shi, N., J. Xu, S.W. Wurster, H. Guo, J. Woodring, L.P. Van Roekel, and H.W. Shen. 2022. "GNN-Surrogate: A Hierarchical and Adaptive Graph Neural Network for Parameter Space Exploration of Unstructured-Mesh Ocean Simulations." *arXiv*:2202.08956. https://ui.adsabs.harvard.edu/abs/2022arXiv220208956S.

Talirz, L., S. Kumbhar, E. Passaro, A.V. Yakutovich, V. Granata, F. Gargiulo, M. Borelli, et al. 2020. "Materials Cloud, a Platform for Open Computational Science." *Scientific Data* 7(1):299.

Teranishi, K., H. Menon, W.F. Godoy, P. Balaprakash, D. Bau, T. Ben-Nun, A. Bhatele, et al. 2025. "Leveraging AI for Productive and Trustworthy HPC Software: Challenges and Research Directions." *arXiv*. https://doi.org/10.48550/arXiv.2505.08135.

Tian, D., J. Deng, E. Zio, F. Maio, and F. Liao. 2018. "Failure Modes Detection of Nuclear Systems Using Machine Learning." In *2018 5th International Conference on Dependable Systems and Their Applications (DSA)*. IEEE. https://doi.org/10.1109/DSA.2018.00017.

Tidjon, L.N., and F. Khomh. 2022. "The Different Faces of AI Ethics Across the World: A Principle-Implementation Gap Analysis." *arXiv*:2206.03225. https://ui.adsabs.harvard.edu/abs/2022arXiv220603225N.

Wang, S., J. González-Cao, H. Islam, M. Gómez-Gesteira, and C. Guedes Soares. 2022. "Uncertainty Estimation of Mesh-Free and Mesh-Based Simulations of the Dynamics of Floaters." *Ocean Engineering* 256:111386.

Wang, Z., M. Daeipour, and H. Xu. 2023. "Quantification and Propagation of Aleatoric Uncertainties in Topological Structures." *Reliability Engineering and System Safety* 233:109122.

Wang, Z., H. Wei, R. Tian, and S. Tan. 2025. "A Review of Data-Driven Fault Diagnosis Method for Nuclear Power Plant." *Progress in Nuclear Energy* 186:105785.

Wong, M.L., C.E. Cleland, D. Arend, S. Bartlett, H.J. Cleaves, H. Demarest, A. Prabhu, J.I. Lunine, and R.M. Hazen. 2023. "On the Roles of Function and Selection in Evolving Systems." *Proceedings of the National Academy of Sciences of the United States of America* 120(43):e2310223120.

Xu, Z., J. Ren, Y. Zhang, J.M. Gonzalez Ondina, M. Olabarrieta, T. Xiao, W. He, et al. 2024. "Accelerate Coastal Ocean Circulation Model with AI Surrogate." *arXiv*:2410.14952. https://ui.adsabs.harvard.edu/abs/2024arXiv241014952X.

Yang, Y., W. Youyou, and B. Uzzi. 2020. "Estimating the Deep Replicability of Scientific Findings Using Human and Artificial Intelligence." *Proceedings of the National Academy of Sciences of the United States of America* 117(20):10762–10768.

Yarger, D., B.M. Wagman, K. Chowdhary, and L. Shand. 2024. "Autocalibration of the E3SM Version 2 Atmosphere Model Using a PCA-Based Surrogate for Spatial Fields." *Journal of Advances in Modeling Earth Systems* 16:e2023MS003961.

Yuan, E.C.-Y., Y., Liu, J. Chen, P. Zhong, S. Raja, T. Kreiman, S. Vargas, et al. 2025. "Foundation Models for Atomistic Simulation of Chemistry and Materials." *arXiv Physics*. https://doi.org/10.48550/arXiv.2503.10538.

Zhang, Z. 2024. "MODNO: Multi-Operator Learning with Distributed Neural Operators." *Computer Methods in Applied Mechanics and Engineering* 431:117229.

A

Statement of Task

A National Academies of Sciences, Engineering, and Medicine consensus study will assess the state of the art in foundation models and their use across science research domains relevant to the Department of Energy mission. The study will address the following questions:

- What are some exemplar use cases where foundation models could impact scientific discovery and innovation?
- How can foundation models be used in conjunction with traditional modeling, computational, and data science approaches?
- How can challenges such as verification, validation, uncertainty quantification, and reproducibility best be addressed to advance trustworthy foundation models?
- What are priority research areas for investments to advance the development and use of foundation models in scientific applications? What are the trade-offs in investing in foundation models versus other mathematical and computational approaches?

B

Public Meeting Agendas

February 11, 2025 [4 pm–6 pm ET]

VIRTUAL OPEN SESSION

[4 pm] Welcome and Introductions
- Dona Crawford, Committee Chair

[4:10 pm] Introduction from Department of Energy (DOE) Advanced Scientific Computing Research (ASCR)
- Hal Finkel, Director Computational Science Research and Partnerships Division, ASCR
- Steven Lee, Program Manager for Applied Mathematics and AI

[4:50 pm] Introduction from DOE National Nuclear Security Administration
- Si Hammond, Federal Program Manager
- Thuc Hoang, Deputy Assistant Deputy Administrator for Advanced Simulation and Computing

[5:30 pm] Questions from the Committee
- Dona Crawford, Committee Chair

[6 pm] Committee Closed Session

- Committee Reactions

March 6, 2025 [9 am–5 pm PT]
Location: Board Room, Beckman Center, 100 Academy Way, Irvine, CA 92617

HYBRID OPEN SESSION

[9 am] Welcome and Introductions

- Dona Crawford, Committee Chair

[9:20 am] Industry Perspective
20-minute presentation from each speaker with 35-minute Q&A for the group
Petros Koumoutsakos, Moderator

- Vivek Natarajan, Google DeepMind
- Sebastian Nowozin, Google DeepMind

Panel Q&A, moderated by Petros Koumoutsakos

[10:35 am] Break

[10:50 am] DOE National Lab Panel 1
20-minute presentation from each speaker with 30-minute Q&A for the group
Syed Bahauddin Alam, Moderator

- Earl Lawrence, Los Alamos National Laboratory
- Michael Mahoney, Lawrence Berkeley National Laboratory
- Chris Ritter, Idaho National Laboratory

Panel Q&A, moderated by Syed Bahauddin Alam

[12:20 pm] Lunch

[1:20 pm] DOE National Lab 2
20-minute presentation from each speaker with 40-minute Q&A for the group
Dan Meiron, Moderator

- Hendrik Hamann, Brookhaven National Laboratory
- Rick Stevens, Argonne National Laboratory
- Georgia Tourassi, Oak Ridge National Laboratory

Panel Q&A, moderated by Dan Meiron

[3:00 pm] Break

[3:20 pm] Applications and Foundation Model Users Panel 1
20-minute presentation from each speaker with 40-minute Q&A for the group
Krishna Garikipati, Moderator

- Rémi Lam, Google DeepMind
- James Warren, National Institute of Standards and Technology
- Bin Yu, University of California, Berkeley

Panel Q&A, moderated by Krishna Garikipati

[5:00 pm] Open Session Adjourn

March 7, 2025 [9 am–3 pm PT]
Location: Board Room, Beckman Center, 100 Academy Way, Irvine, CA 92617

HYBRID OPEN SESSION

[9 am] Welcome Back and Introductions

- Dona Crawford, Committee Chair

[9:20 am] Applications and Foundation Model Users Panel 2
20-minute presentation from each speaker with 35-minute Q&A for the group
Marta D'Elia, Moderator

- William Collins, Lawrence Berkeley National Laboratory
- Ann Speed, Sandia National Laboratories

Panel Q&A, moderated by Marta D'Elia

[10:35 am] Break

[10:50 am] DOE National Lab Panel 3
20-minute presentation from each speaker with 30-minute Q&A for the group
Brian Kulis, Moderator

- Kevin Dixon, Sandia National Laboratories
- Kelly Rose, National Energy Technology Laboratory
- Brian Spears, Lawrence Livermore National Laboratory

Panel Q&A moderated by Brian Kulis

[12:20 pm] Lunch—Open Session Adjourns

[1:20 pm] Closed Committee Session

May 6, 2025 [1 pm–3 pm ET]

VIRTUAL OPEN SESSION

[1 pm] Welcome and Introductions
- Dona Crawford, Committee Chair

[1:15 pm] Foundation Model Presentation
- Omar Ghattas, University of Texas at Austin

[1:45 pm] Committee Closed Session—Open Meeting Adjourns

[3:00 pm] Closed Meeting Adjourns

May 20, 2025 [4 pm–6 pm ET]

VIRTUAL OPEN SESSION

[4 pm] Welcome and Introductions
- Dona Crawford, Committee Chair

[4:15 pm] Foundation Model Presentation
- Tzanio Kolev, Lawrence Livermore National Laboratory

[4:45 pm] Committee Closed Session—Open Meeting Adjourns

[6:00 pm] Closed Meeting Adjourn

C

Acronyms and Abbreviations

AI	artificial intelligence
AM	advanced manufacturing
ASCR	Advanced Scientific Computing Research
DOE	Department of Energy
ECP	Exascale Computing Project
FFRDC	federally funded research and development center
FNO	Fourier neural operator
GPT	generative pretrained transformer
HPC	high-performance computing
LDRD	laboratory-directed research and development
LLM	large language model
LLNL	Lawrence Livermore National Laboratory
MoE	mixture-of-experts
NNSA	National Nuclear Security Administration
NREL	National Renewable Energy Laboratory
PDE	partial differential equations
PNNL	Pacific Northwest National Laboratory

RAG	retrieval-augmented generation
SSP	U.S. Stockpile Stewardship Program
VVUQ	verification, validation, and uncertainty qualification

D

Committee Member Biographical Information

DONA L. CRAWFORD, *Chair*, retired as the associate director for computation from the Lawrence Livermore National Laboratory (LLNL), where she led the laboratory's high-performance computing efforts. In that capacity, Crawford was responsible for the development and deployment of an integrated computing environment for petascale simulations of complex physical phenomena. Prior to her LLNL appointment in 2001, Crawford was with Sandia National Laboratories since 1976 serving on many leadership projects including the Accelerated Strategic Computing Initiative and the Nuclear Weapons Strategic Business Unit. Crawford serves on the National Academies of Sciences, Engineering, and Medicine's Laboratory Assessments Board and has previously served on several National Academies' committees including the Committee to Evaluate Post-Exascale Computing for the National Nuclear Security Administration, the Committee to Review Governance Reform in the National Nuclear Security Administration, and the Committee to Evaluate the National Science Foundation's Vertically Integrated Grants for Research and Education Program. She received her MS in operations research from Stanford University.

SYED BAHAUDDIN ALAM is an assistant professor of nuclear, plasma, and radiological engineering at the University of Illinois Urbana-Champaign (UIUC), where he leads the MARTIANS (Machine Learning & ARTificial Intelligence for Advancing Nuclear Systems) Laboratory. He was named as the national artificial intelligence (AI) leader in UIUC's official response to the White House AI Action Plan (2025). He holds a joint appointment at the National Center for Supercomputing Applications. Alam's research expertise centers on energy-efficient AI and digital twins, with a primary focus on developing real-time AI algorithms for nuclear and energy systems. He has been recognized with numerous prestigious awards, including the Nuclear News 40 Under 40, Dean's Award for Excellence in Research from the UIUC Grainger College of Engineering, Illinois Innovation Award finalist for excellence in cutting-edge innovation, a "Top of Minds" feature by UIUC Grainger College, the Cambridge Philosophical Society Award, the American Nuclear Society Best Paper Award, the Cambridge Trust Award, and an Outstanding Teaching Award. He earned his PhD (2018) and MPhil (2013) in nuclear engineering from the University of Cambridge and BSc (2011) in electrical and electronic engineering from the Bangladesh University of Engineering and Technology.

MARTA D'ELIA is the director of AI and ModSim at Atomic Machines and an adjunct professor at the Stanford University Institute for Computational & Mathematical Engineering. She previously worked at Pasteur Labs, Meta, and Sandia National Laboratories as a principal scientist and tech lead. She holds a PhD in applied mathematics and master's and bachelor's degrees in mathematical engineering. Her work deals with design and analysis of machine learning (ML) models and optimal design and control for complex industrial applications. She is an expert in nonlocal modeling and simulation, optimal control, and scientific ML. She is an associate editor of *Society and Industrial and Applied Mathematics* (SIAM) and *Nature* journals, a member of the SIAM industry committee, the vice chair of the SIAM Northern California section, and a member of the NVIDIA advisory board for scientific ML.

KRISHNA GARIKIPATI obtained his PhD at Stanford University in 1996, and after a few years of postdoctoral work, he joined the University of Michigan in 2000, rising to professor in the Departments of Mechanical Engineering and Mathematics. Between 2016 and 2022, he served as the director of the Michigan Institute for Computational Discovery & Engineering. In January 2024 he moved to the Department of Aerospace and Mechanical Engineering at the University of Southern California. His research is in computational science, with applications drawn from biophysics, materials physics, mechanics, and mathematical biology. Of recent interest are data-driven approaches to computational science. He has been awarded the Department of Energy Early Career Award for Scientists and Engineers, the Presidential Early Career Award for Scientists and Engineers, and a Humboldt Research Fellowship. He is a fellow of the U.S. Association for Computational Mechanics, the International Association for Computational Mechanics, and the Society of Engineering Science; a Life Member of Clare Hall at the University of Cambridge; and a visiting scholar in computational biology at the Flatiron Institute of the Simons Foundation.

SHIRLEY HO is a senior research scientist at the Center for Computational Astrophysics at the Simons Foundation. She joined the Foundation in 2018 to lead the Cosmology X Data Science group. Her research interests range from cosmology to developing new ML methods for scientific data that leverage shared concepts across scientific domains. Ho has extensive expertise in astrophysical theory, observation, and data science. She focuses on novel statistical and ML tools to address cosmic mysteries such as the origins and fate of the universe. Ho analyzes data from surveys by the Atacama Cosmology Telescope, the Euclid Observatory, the Large Synoptic Survey Telescope, the Simons Observatory, the Sloan Digital Sky Survey, and the Roman Space Telescope, among others, to understand our universe's evolution. She earned her PhD in astrophysical sciences from Princeton University in 2008 and BS degrees in computer science and physics from University of California, Berkeley, in 2004. Ho was previously a Chamberlain and Seaborg Fellow at Lawrence Berkeley National Laboratory (LBNL). She joined Carnegie Mellon University as an assistant professor in 2011, becoming the Cooper Siegel Career Development Chair Professor and a tenured associate professor. In 2016 she moved to LBNL as a senior scientist.

SCOTT H. HOLAN is a Curators' Distinguished Professor and the department chair in the Department of Statistics and Data Science at the University of Missouri and serves as a senior research fellow in the Research and Methodology Directorate at the U.S. Census Bureau. His research expertise includes developing statistical and ML methodology for dependent data (spatial, spatiotemporal, functional, and multivariate, among others), Bayesian methods, environmental and ecological statistics, official statistics, and survey methodology. He is an elected Fellow of the American Statistical Association (2014), an elected member of the International Statistical Institute (2017), an elected Fellow of the Institute of Mathematical Statistics (2021), and an elected Fellow of the American Association for the Advancement of Science (2024). Holan was a previous co-awardee of the Statistical Partnerships Among Academe, Industry, and Government Award (2017).

MICHAEL KEARNS is a professor and the National Center chair of the Department of Computer and Information Science at the University of Pennsylvania and the founding director of the Warren Center for Network and Data Sciences. His research interests include topics in ML, AI, algorithmic game theory and microeconomics, computational social science, and quantitative finance and algorithmic trading. Kearns often examines problems in these areas using methods and models from theoretical computer science and related disciplines. He also often participates in empirical and experimental projects, including applications of ML to problems in algorithmic trading and quantitative finance, and human-subject experiments on strategic and economic interaction in social networks.

Kearns spent 1991–2001 in ML and AI research at AT&T Bell Labs and in the last 4 years of his appointment was head of the AI department, which conducted a broad range of systems and foundational AI work. Kearns received his undergraduate degrees from the University of California, Berkeley, in mathematics and computer science and his PhD in computer science from Harvard University. In 2020, Kearns joined Amazon Web Services as an Amazon Scholar, focusing on fairness, privacy, and other "responsible AI" topics. He is an elected member of the National Academy of Sciences.

PETROS KOUMOUTSAKOS is the Herbert S. Winokur Jr. Professor for Computing in Science and Engineering. He also currently holds a visiting researcher position at Google DeepMind in London. He studied Naval Architecture (diploma from the National Technical University of Athens, MEng from the University of Michigan, and received a PhD in aeronautics and applied mathematics from the California Institute of Technology [Caltech]). He has conducted postdoctoral studies at the Center for Parallel Computing at Caltech and at the Center for Turbulent Research at Stanford University and NASA Ames. He has served as the chair of computational science at ETHZurich (1997–2020). Koumoutsakos is an elected Fellow of the American Society of Mechanical Engineers, the American Physical Society, and the Society of Industrial and Applied Mathematics. He is a recipient of the Advanced Investigator Award from the European Research Council and the Association for Computing Machinery's Gordon Bell prize in supercomputing. He is an elected International Member of the National Academy of Engineering.

BRIAN KULIS is an associate professor at Boston University, with appointments in the Department of Electrical and Computer Engineering, the Department of Computer Science, the Faculty of Computing and Data Sciences, and the Division of Systems Engineering. From 2019 to 2023, he was also an Amazon Scholar, working with the Alexa team. Previously, he was the Peter J. Levine Career Development Assistant Professor at Boston University. Before joining Boston University, he was an assistant professor in computer science and in statistics at Ohio State University. Prior to that he was a postdoctoral fellow at the University of California, Berkeley, Electrical Engineering & Computer Sciences. His research focuses on ML, statistics, computer vision, and large-scale optimization. He obtained his PhD in computer science from the University of Texas in 2008 and his BA from Cornell University in computer science and mathematics in 2003. For his research, he has won three best paper awards at top-tier conferences—two at the International Conference on Machine Learning (2005 and 2007) and one at the IEEE Conference on Computer Vision and Pattern Recognition (2008). He was also the recipient of a National Science Foundation (NSF) CAREER Award in 2015.

DANIEL I. MEIRON is currently a professor of aerospace and applied and computational mathematics. His research interests are primarily in computational fluid dynamics with connections to high-performance computing. He also has interests in computational materials science. He received an ScD in applied mathematics at the Massachusetts Institute of Technology working under Steven A. Orszag. He has participated as part of a recent National Academies' study on exascale computing.

NATHANIEL TRASK recently joined the Department of Mechanical Engineering and Applied Mechanics at the University of Pennsylvania after spending 8 years as technical staff at Sandia National Laboratories. His research focuses on developing foundational aspects of scientific machine learning (SciML) for high-consequence engineering settings. By integrating concepts from modern physics and probability into the design of deep learning architectures, he leads a research program employing SciML for scientific discovery as well as to construct digital twins of complex systems. He is the deputy director of the Scalable, Efficient and Accelerated Causal Reasoning Operators, Graphs and Spikes for Earth and Embedded Systems Center, an Office of Science funded multi-institutional center developing next-generation AI architectures for Earth and embedded systems. He has received the Department of Energy Early Career Award, as well as the NSF Mathematical Science Postdoctoral Fellowship. His doctoral training was in applied mathematics, with a focus on developing novel optimization-based discretizations of partial differential equations to simulate multiphysics and multiscale problems. After moving to Sandia National Laboratories for a fellowship, he went on to work extensively on ML applied to material science and physics in extreme environments.